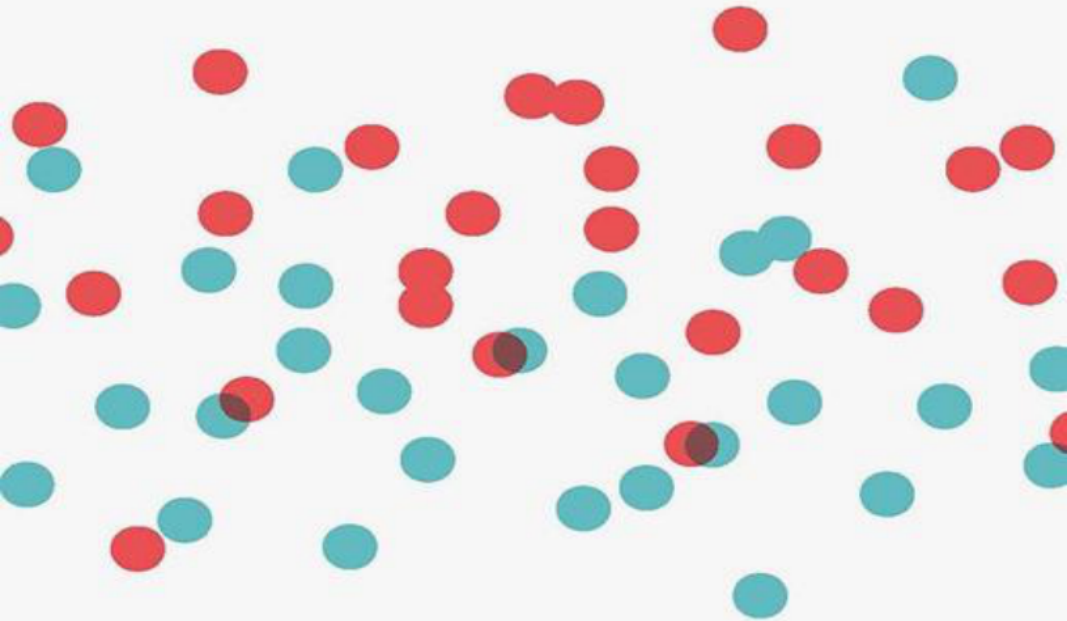
David Spiegelhalter

El arte de la estadística

Cómo aprender de los datos



Capitán Swing

El arte de la estadística

Cómo aprender de los datos

DAVID SPIEGELHALTER

Traducción de

Francisco Herreros

Capitán Swing

LISTA DE FIGURAS

0.1 Edad y año de la muerte de las víctimas de Harold Shipman

0.2 Hora de la muerte de los pacientes de Harold Shipman

0.3 Ciclo PPDAC de resolución de problemas

1.1 Tasas de supervivencia a los 30 días de cirugía cardiaca

1.2 Proporción de operaciones de corazón a niños por hospital

1.3 Porcentaje de operaciones de corazón a niños por hospital

1.4 Riesgo asociado a comer bocadillos de beicon

2.1 Tarro de gominolas

2.2 Diferentes formas de presentar estimaciones sobre gominolas

2.3 Estimaciones sobre gominolas presentadas en un diagrama con escala logarítmica

2.4 Número declarado de parejas sexuales de sexo opuesto a lo largo de la vida

2.5 Tasas de supervivencia cruzadas con el número de operaciones de cirugía cardiaca en niños

2.6 Coeficientes de correlación de Pearson de 0

2.7 Tendencias demográficas mundiales

2.8 Incremento demográfico relativo por país

2.9 Popularidad del nombre «David» a lo largo del tiempo

2.10 Infografía sobre actitudes y estilos de vida sexuales

3.1 Diagrama de inferencia inductiva

3.2 Distribución del peso al nacer

5.1 Diagrama de dispersión de la altura de los hijos frente a la de los padres

5.2 Modelo de regresión logística para datos de cirugía cardiaca en niños

6.1 Lápida de una víctima del Titanic

6.2 Estadísticas de resumen sobre supervivencia para los pasajeros del Titanic

6.3 Árbol de clasificación de datos del Titanic

6.4 Curvas ROC para algoritmos aplicados a conjuntos de entrenamiento y de validación

6.5 Probabilidad de sobrevivir al hundimiento del Titanic

6.6 Árbol de clasificación sobreajustado para los datos del Titanic

6.7 Tasas de supervivencia posoperatoria para mujeres con cáncer de mama

7.1 Distribución empírica del número de parejas sexuales para varios tamaños muestrales

7.2 Re-muestreos bootstrap a partir de una muestra original de 50

7.3 Distribución bootstrap de medias para varios tamaños muestrales

7.4 Regresión bootstrap con datos de Galton madre-hija

8.1 Simulación de los juegos del Chevalier de Méré

8.2 Árbol de frecuencia esperada para dos lanzamientos de moneda

8.3 Árbol de probabilidad para lanzamientos de dos monedas

8.4 Árbol de frecuencia esperada para revisiones de cáncer de mama

8.5 Número de homicidios observado y esperado

9.1 Distribución de probabilidad de personas zurdas

9.2 Gráfico de embudo de tasas de muerte por cáncer colorrectal

9.3 Gráfico de la BBC de encuestas de opinión previas a las elecciones generales de 2017

9.4 Tasas de homicidios en Inglaterra y Gales

10.1 Razón de sexos para bautismos en Londres, 1629-1710

10.2 Distribución empírica de las diferencias observadas en la

proporción

de personas que se cruzan de brazos con el izquierdo/derecho encima

10.3 Número acumulado de certificados de defunción firmados por Shipman

10.4 Prueba de razón de probabilidad secuencial para la detección de una duplicación de la tasa de mortalidad

10.5 Frecuencias esperadas de los resultados de 1.000 pruebas de hipótesis

11.1 Árbol de frecuencia esperada para el problema de las tres monedas

11.2 Árbol de frecuencia esperada para el dopaje en el deporte

11.3 Árbol de frecuencia esperada inversa para el dopaje en el deporte

11.4 Mesa de «billar» de Bayes

12.1 Flujos tradicionales de información para la evidencia estadística

LISTA DE TABLAS

1.1 Resultados de la cirugía cardiaca infantil

1.2 Métodos para comunicar el riesgo de cáncer colorrectal a lo largo de la vida en consumidores de beicon

2.1 Estadísticos descriptivos para estimaciones sobre gominolas

2.2 Estadísticos descriptivos para el número de parejas sexuales a lo largo de la vida

4.1 Resultados para pacientes del Estudio de Protección Cardiaca

4.2 Ilustración de la paradoja de Simpson

5.1 Estadísticos de resumen de la estatura de padres y sus hijos adultos

5.2 Correlación entre las estaturas de hijos adultos y padres del mismo género

5.3 Resultados de una regresión lineal múltiple que relaciona la estatura de los hijos adultos con la de la madre y el padre

6.1 Matriz de errores del árbol de clasificación sobre datos de entrenamiento y validación del Titanic

6.2 Predicción ficticia de «probabilidad de precipitación»

6.3 Resultados de una regresión logística con datos de supervivientes del Titanic

6.4 Desempeño de distintos algoritmos sobre datos de prueba del Titanic 6.5 Tasas de supervivencia al cáncer de mama con el algoritmo Predict 2.1 7.1 Estadísticos de resumen de parejas sexuales a lo largo de la vida declaradas por hombres

7.2 Medias muestrales de parejas sexuales a lo largo de la vida declaradas por hombres

9.1 Comparación de intervalos de confianza exactos y bootstrap

10.1 Contingencia del cruzado de brazos por género

10.2 Recuento observado y esperado de cruzado de brazos por género

10.3 Días observados y esperados con cada número de incidentes de

homicidio

10.4 Resultados del Estudio de Protección Cardiaca con intervalos de confianza y valores P

10.5 Resultado en R de una regresión múltiple con datos de Galton

10.6 Resultados posibles de una prueba de hipótesis

11.1 Razones de verosimilitud para la evidencia sobre el esqueleto de Ricardo III

11.2 Interpretaciones verbales recomendadas de las razones de verosimilitud

11.3 Escala de Kass y Raftery para las interpretaciones de factores de Bayes

13.1 Predicciones de la encuesta a pie de urna para tres elecciones generales recientes

Agradecimientos

Detrás de todos los conocimientos adquiridos en mi larga carrera en el campo de la estadística hay un buen número de estimulantes colegas de los que he aprendido. Son demasiados para que los cuente incluso un estadístico como yo, pero una lista breve de aquellos a los que he arrebatado más conocimientos incluiría a Nicky Best, Sheila Bird, David Cox, Philip Dawid, Stephen Evans, Andrew Gelman, Tim Harford, Kevin McConway, Wayne Oldford, Sylvia Richardson, Hetan Shah, Adrian Smith y Chris Wild. Les estoy muy agradecido a ellos y a tantos otros por animarme con un tema tan complejo.

El desarrollo del libro ha sido más bien lento, debido por entero a mi tendencia crónica a dejar todo para el final. Por ello, me gustaría agradecer en primer lugar a Laura Stickney, de Penguin, no solo por encargarme el libro, sino por mantener la calma a medida que pasaban los meses (y los años), e incluso cuando, ya terminado el libro, todavía éramos incapaces de llegar a un acuerdo sobre el título. Y todo mi agradecimiento a Jonathan Pegg por negociarme un acuerdo excelente, a Jane Birdsell por demostrar una paciencia enorme editando el manuscrito y a todo el personal de producción de Penguin por su meticuloso trabajo.

Estoy muy agradecido por los permisos concedidos para adaptar ilustraciones, en concreto a Chris Wild (figura 0.3), James Grime (figura 2.1), Cath Mercer, de Natsal (figuras 2.4 y 2.10), la Oficina Nacional de Estadística (figuras 2.9, 8.5 y 9.4), el Departamento de Salud Pública de Inglaterra (figura 6.7), Paul Barden (figura 9.2) y a la BBC (figura 9.3). El uso de información del sector público británico está autorizado por la Autorización de Gobierno Abierto v3.0.

No soy un buen programador de R, así que recurrí a la valiosa ayuda de Matthew Pearce y Maria Skoularidou para los análisis y los gráficos. También me cuesta escribir, por lo que estoy en deuda con muchas personas que leyeron y comentaron varios capítulos, por ejemplo George Farmer, Alex Freeman, Cameron Brick, Michael Posner, Sander van der Linden y Simone Warr; en particular, Julian Gilbey demostró tener un excelente ojo para detectar errores y ambigüedades.

Por encima de todo, debo agradecer a Kate Bull no solo por sus comentarios vitales al texto, sino también por apoyarme en épocas tanto buenas (de escritura en una cabaña en la playa de Goa) como no tan buenas (un lluvioso febrero haciendo malabarismos con

demasiados compromisos).

También estoy profundamente agradecido a David y Claudia Harding tanto por su apoyo financiero como por su inagotable aliento, que me han permitido hacer tantas cosas divertidas en los últimos diez años.

Finalmente, por mucho que quisiera encontrar algún otro chivo expiatorio, me temo que debo aceptar toda la responsabilidad por las inevitables deficiencias del libro.

Código para los ejemplos

El código R y los datos para reproducir la mayor parte de los análisis y las figuras están disponibles en https://github.com/dspiegel29/ArtofStatistics. Agradezco la ayuda recibida para preparar este material.

El arte de la estadística

Cómo aprender de los datos

A los estadísticos de todas partes,
con su entrañable pedantería, generosidad,
integridad y deseo de usar los datos
de la mejor manera posible.

Introducción

«Los números no pueden hablar por sí mismos.

Nosotros hablamos por ellos. Les conferimos significado».

NATE SILVER, La señal y el ruido[1]

Por qué necesitamos la estadística

Harold Shipman fue el asesino convicto más prolífico de Gran Bretaña, aunque no encajaba en el perfil arquetípico del asesino en serie. Médico de familia de maneras suaves, entre 1975 y 1998 inyectó a un mínimo de doscientos quince de sus mayoritariamente ancianos pacientes una sobredosis masiva de opiáceos. Al final, cometió el error de falsificar el testamento de una de sus víctimas para quedarse con parte de su dinero. Su hija, abogada, empezó a sospechar, y un análisis forense del ordenador de Shipman mostró que había modificado el historial de sus pacientes para que sus víctimas apareciesen como más enfermas de lo que realmente estaban. Era un entusiasta precoz de la tecnología, pero, aun así, no era lo suficientemente experto como para saber que cada cambio que hacía quedaba registrado (por cierto, este es un buen ejemplo de cómo los datos pueden revelar significados ocultos).

De sus pacientes que no habían sido incinerados, quince fueron exhumados, en los cuales se encontraron niveles letales de diamorfina, el formato médico de la heroína. Shipman fue procesado por quince asesinatos en 1999, pero decidió no defenderse y no pronunció ni una sola palabra en el juicio. Fue declarado culpable y condenado a cadena perpetua, y se puso en marcha una investigación pública para determinar qué crímenes pudo haber cometido aparte de aquellos por los que había sido juzgado, y si podía haber sido detenido antes. Yo formé parte de un grupo de estadísticos que fueron convocados como testigos expertos en esta investigación, que concluyó que con toda seguridad había asesinado a doscientos quince de sus pacientes, y posiblemente a cuarenta y cinco más.[2]

Este libro se centrará en el empleo de la ciencia estadística[3] para responder al tipo de preguntas que surgen cuando queremos entender mejor el mundo; algunas de estas preguntas se destacarán en un recuadro. Para empezar a entender el comportamiento de Shipman, lo primero que normalmente nos preguntaremos es:

¿A qué tipo de personas asesinó Harold Shipman, y cuándo murieron?

La investigación pública proporcionó detalles sobre la edad, el género y la fecha de la muerte de cada una de las víctimas. La figura 0.1 es una visualización bastante sofisticada de los datos. Muestra un diagrama de dispersión de la edad de la víctima cruzada con su fecha de fallecimiento, y la intensidad de los puntos indica si la víctima era hombre o mujer. Sobre los ejes se han superpuesto diagramas de barras que muestran las pautas de edad (en intervalos de cinco años) y los años.

Se pueden extraer algunas conclusiones simplemente mirando la figura. Hay más puntos negros que blancos, de manera que las víctimas de Shipman fueron mayoritariamente mujeres. El diagrama de barras a la derecha de la figura muestra que la mayor parte de sus víctimas eran septuagenarias u octogenarias, pero un vistazo a los puntos nos revela que, aunque inicialmente todos eran ancianos, a medida que pasan los años aparecen algunos casos más jóvenes. El diagrama de barras en la parte superior de la figura muestra claramente un hiato alrededor de 1992, en el que no se produjeron asesinatos. Al parecer, antes de esa fecha Shipman había trabajado en una consulta compartida con otros doctores, pero a partir de entonces, probablemente al sentirse bajo sospecha, montó su propia consulta médica. Después de esto, sus actividades se aceleraron, como demuestra el diagrama de barras superior.

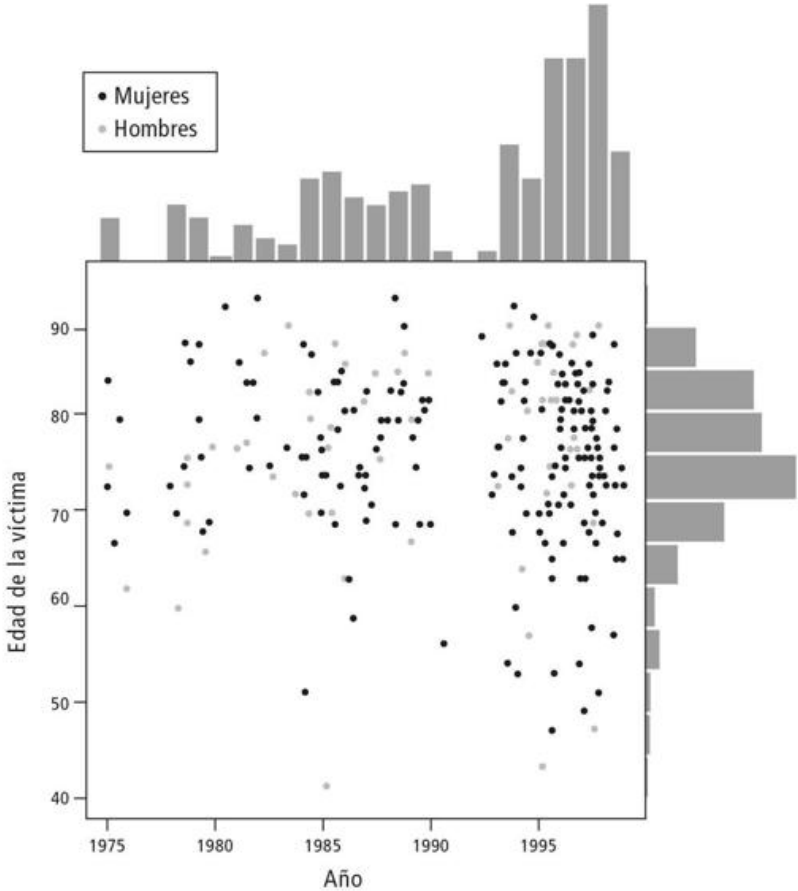


Figura 0.1. Diagrama de dispersión que muestra la edad y el año de la muerte de las doscientas quince víctimas confirmadas de Harold Shipman. Se han añadido diagramas de barras sobre los ejes para mostrar las pautas de edades y los años en los que cometió los asesinatos.

Este análisis de las víctimas identificadas por la investigación plantea otras preguntas sobre la forma en la que cometió sus asesinatos. Hay alguna evidencia estadística derivada de los datos sobre el momento de la muerte de sus supuestas víctimas, recogido en el certificado de defunción. La figura 0.2 es un gráfico lineal que compara las horas en las que murieron los pacientes de Shipman con las horas en las que murieron una muestra de pacientes de otros médicos de familia locales. La pauta no requiere un análisis especialmente sutil: la conclusión se conoce normalmente como «interocular», dado que parece que te entra por los ojos. Los pacientes de Shipman solían morir abrumadoramente en las primeras horas de la tarde.

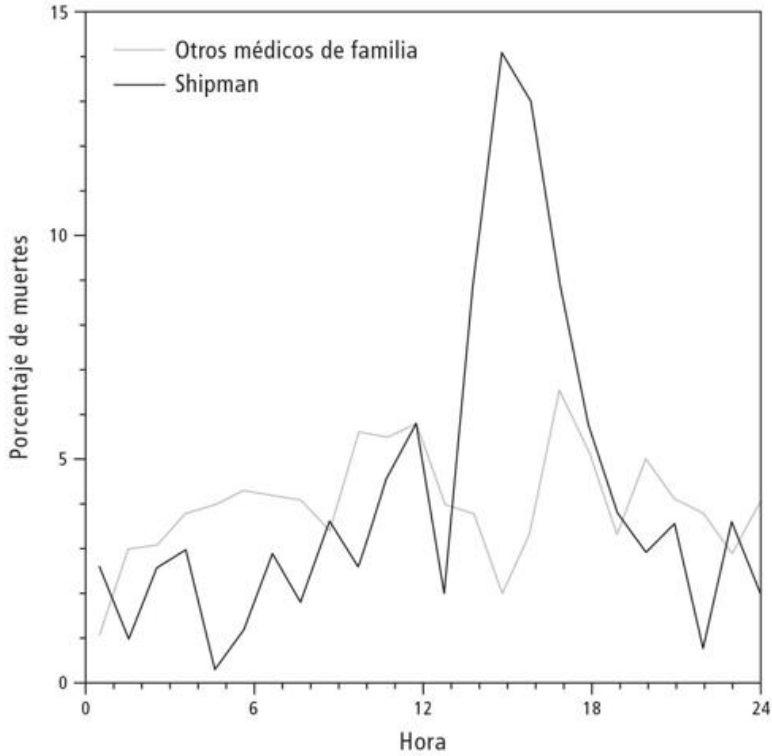


Figura 0.2. Hora de la muerte de los pacientes de Harold Shipman, comparada con las horas en las que murieron los pacientes de otros médicos de familia locales. La pauta no requiere un análisis estadístico sofisticado.

Los datos no nos dicen por qué tendían a morir en esos precisos momentos, pero una investigación subsiguiente reveló que llevaba a cabo sus visitas a domicilio después de comer, cuando estaba generalmente a solas con sus pacientes ancianos. Les ofrecía una inyección que supuestamente los haría sentir más cómodos, pero que, de hecho, era una dosis letal de diamorfina: después de que un paciente muriese pacíficamente delante de él, cambiaba su historial médico para que pareciese que se trataba de una muerte natural esperada. La magistrada Janet Smith, que encabezaba la investigación pública, dijo posteriormente: «Todavía siento que es indescriptiblemente horrible, simplemente indescriptible, impensable e inimaginable que día tras día fingiese ser ese doctor maravillosamente bondadoso mientras llevaba en su maletín su arma letal […] que sacaba de la manera más trivial del mundo».

Estaba asumiendo un riesgo, dado que una simple autopsia le habría

descubierto, pero dada la edad de sus pacientes y las causas aparentemente naturales de sus muertes, no se llevó a cabo ninguna. Y sus razones para cometer estos asesinatos nunca han sido explicadas: no habló en su juicio, nunca comentó sus fechorías a nadie, incluida su familia, y se suicidó en la cárcel, convenientemente justo a tiempo para que su mujer recibiese su pensión.

Podemos pensar en este tipo de trabajo iterativo y exploratorio como una forma de estadística «forense», algo que en este caso era literalmente correcto. No hay matemáticas ni teoría, solo una búsqueda de pautas que podrían llevar a preguntas más interesantes. Los detalles de las fechorías de Shipman fueron establecidos mediante evidencia específica de cada caso particular, pero este análisis de datos era el sostén de una comprensión más general de cómo llevó a cabo sus crímenes.

Más adelante, en el capítulo 10, veremos si el análisis estadístico formal podría haber ayudado a atrapar antes a Shipman.[4] Mientras tanto, la historia de Shipman demuestra ampliamente el gran potencial que ofrecen los datos para ayudarnos a comprender el mundo y hacer mejores juicios. Esta es la esencia de la ciencia estadística.

Transformar el mundo en datos

El enfoque estadístico en los crímenes de Harold Shipman nos exigía tomar cierta distancia de la larga lista de tragedias individuales de las que era responsable. Todos esos detalles personales únicos, de esas vidas y de sus muertes, tenían que ser reducidos a un conjunto de hechos y números que pudieran ser contados y presentados en gráficos. A primera vista, esto podría parecer frío y deshumanizador, pero si usamos la ciencia estadística para iluminar el mundo, entonces nuestras experiencias diarias tienen que convertirse en datos, y esto significa categorizar y clasificar eventos, registrar mediciones, analizar los resultados y comunicar las conclusiones.

Algo aparentemente tan simple como categorizar y clasificar puede, no obstante, presentar un serio desafío. Tomemos, por ejemplo, la siguiente pregunta básica, que debería resultar de interés para cualquier persona preocupada por el medio ambiente:

¿Cuántos árboles hay en el planeta?

Antes incluso de comenzar a pensar en cómo deberíamos contestar a esta pregunta, tenemos primero que dejar sentada una cuestión

bastante básica. ¿Qué es un «árbol»? Podría usted pensar que sabe distinguir lo que es un árbol con solo ver uno, pero su juicio podría diferir sustancialmente del de otras personas que podrían considerarlo más bien un arbusto o un matorral. De manera que, para transformar la experiencia en datos, tenemos que empezar con definiciones rigurosas.

El caso es que la definición oficial de «árbol» es una planta con un tronco de madera que tiene un diámetro suficientemente grande a la altura del pecho, conocido como DAP. El Servicio Forestal de Estados Unidos exige que una planta tenga un DAP mayor de 5 pulgadas (12,7 centímetros) para que sea oficialmente declarada como árbol, pero la mayor parte de las autoridades usan un DAP de 10 centímetros (4 pulgadas).

No obstante, no podemos recorrer el planeta midiendo cada planta con tronco de madera y contando todas las que cumplen este criterio. De manera que los investigadores que estudiaron esta cuestión adoptaron un enfoque más pragmático: primero se centraron en una serie de áreas con un tipo de paisaje común, lo que se conoce como bioma, y contaron el número medio de árboles por kilómetro cuadrado. Después usaron imágenes de satélite para estimar el área total del planeta cubierta por cada tipo de bioma, llevaron a cabo una cierta modelización estadística compleja, y, finalmente, generaron una estimación total de 3,04 billones (es decir, 3.040.000.000.000) de árboles en el planeta. Parecen muchísimos, hasta que nos enteramos de que los investigadores piensan que llegó a haber el doble.[5]

Si las autoridades difieren acerca de qué es un árbol, no debería sorprendernos que conceptos aún más nebulosos sean más difíciles de definir. Por tomar un ejemplo extremo, la definición oficial de «desempleo» en el Reino Unido ha cambiado al menos treinta y una veces entre 1979 y 1996.[6] La definición de producto interior bruto (PIB) se revisa continuamente, como por ejemplo cuando el comercio de drogas ilegales y la prostitución fueron añadidos al PIB del Reino Unido en 2014; las estimaciones se hicieron usando algunas fuentes de datos inusuales, por ejemplo, Punternet, una página web donde se evaluaban servicios de prostitución, se usó como fuente para los precios de diferentes actividades.[7]

Incluso nuestros sentimientos más personales pueden ser codificados y sometidos a análisis estadístico. En el año anterior a septiembre de 2017, 150.000 personas en el Reino Unido fueron entrevistadas para una encuesta, y se les hizo la siguiente pregunta: «En líneas generales, ¿cómo de feliz se sintió usted ayer?».[8] La respuesta media, en una

escala de 0 a 10, fue 7,5, una mejora con respecto al 7,3 de 2012, que podría estar relacionada con la recuperación económica desde la crisis financiera de 2008. Las puntuaciones más bajas correspondieron a aquellos cuyas edades estaban comprendidas entre los 50 y los 54, y las más altas entre las personas de 70 a 74 años, una pauta común en el Reino Unido.[9]

Medir la felicidad es difícil, pero decidir cuándo alguien está vivo o muerto debería ser mucho más sencillo: como demostrarán los ejemplos de este libro, la supervivencia y la mortalidad son preocupaciones comunes de la ciencia estadística. Pero en Estados Unidos cada estado puede tener su propia definición legal de la muerte, y aunque en 1981 se aprobó la Ley de Uniformidad en la Declaración de Fallecimiento para intentar establecer un modelo común, sigue habiendo pequeñas diferencias. Una persona declarada muerta en Alabama podría, al menos en principio, dejar de estar legalmente muerta si cruzamos la frontera de Florida, donde la certificación debe hacerse por dos médicos cualificados.[10]

Estos ejemplos muestran que las estadísticas siempre se construyen hasta cierto punto sobre la base de juicios, y sería una forma muy obvia de autoengaño el pensar que toda la complejidad de la experiencia personal puede ser codificada inequívocamente y registrada en una hoja de cálculo o en algún otro software. Por difíciles que sean de definir, las medidas y los cálculos de características propias y del mundo que nos rodea son, en definitiva, solamente información, el punto de partida para una compresión real del mundo.

Los datos tienen dos limitaciones principales como fuentes de ese tipo de conocimiento. En primer lugar, siempre son una medida imperfecta de aquello en lo que estamos interesados: el preguntar a un grupo de personas lo felices que eran la semana anterior en una escala de cero a diez difícilmente refleja el bienestar emocional de la nación. En segundo lugar, cualquier cosa que queramos medir diferirá de un sitio a otro, entre una persona y otra, de un momento temporal a otro, y el problema es extraer información significativa de toda esta variabilidad aparentemente aleatoria.

Durante siglos, la ciencia estadística se ha enfrentado a estos desafíos gemelos, y ha jugado un papel principal en los intentos científicos de comprender el mundo. Ha proporcionado la base para interpretar los datos, algo siempre imperfecto, para distinguir importantes relaciones desde la variabilidad de fondo que nos hace únicos a cada uno de nosotros. Pero el mundo siempre está cambiando, a medida que nos

planteamos nuevas preguntas y aparecen nuevas fuentes de datos, y la ciencia estadística, por consiguiente, también tiene que cambiar.

Siempre hemos calculado y medido, pero la estadística moderna comenzó realmente como disciplina en la década de 1650, cuando, como veremos en el capítulo 8, la probabilidad fue entendida correctamente por vez primera gracias a Blaise Pascal y Pierre de Fermat. Una vez que se contaba con esta sólida base matemática para estudiar la variabilidad, el progreso fue extraordinariamente rápido. Cuando se combinó con datos sobre la edad de fallecimiento de la gente, la teoría de la probabilidad proporcionó una base firme para calcular las pensiones y las rentas. La astronomía se transformó radicalmente una vez que los científicos dominaron cómo utilizar la teoría de la probabilidad para manejar la variabilidad de las medidas.

Los más entusiastas entre los victorianos se obsesionaron con reunir datos sobre el cuerpo humano (y sobre todo lo demás) y establecieron una fuerte conexión entre el análisis estadístico y la genética, la biología y la medicina. Más tarde, en el siglo XX, la estadística se hizo más matemática, y, desgraciadamente para muchos estudiantes y profesionales, se convirtió en sinónimo de la aplicación mecánica de una caja de herramientas estadísticas, muchas de ellas con el nombre de unos estadísticos excéntricos y argumentativos con los que nos encontraremos más adelante en este libro.

Esta visión tan común de la estadística como una «caja de herramientas» básicas se enfrenta hoy a varios desafíos. En primer lugar, estamos en la era de la ciencia de los datos, en la que se crean conjuntos de datos complejos a partir de fuentes cotidianas, como los monitores de tráfico, las redes sociales y las compras en internet, y se usan como base para innovaciones tecnológicas tales como la optimización de rutas de viajes, la publicidad focalizada o los sistemas de recomendaciones de compras —nos ocuparemos de los algoritmos basados en los «macrodatos» en el capítulo 6—. Cada vez más, la formación en estadística se ve solamente como uno de los componentes necesarios para ser un científico de datos, junto con conocimientos de gestión de datos, programación y desarrollo de algoritmos, y una gran familiaridad con el objeto de estudio.

Otro desafío a la visión tradicional de la estadística proviene del enorme incremento de la actividad científica, especialmente en las ciencias biomédicas y sociales, combinado con la presión para publicar en revistas de prestigio. Esto ha planteado algunas dudas sobre la credibilidad de parte de la literatura científica, que afirma que muchos «descubrimientos» no pueden ser reproducidos por otros

investigadores —como la disputa eterna sobre si la adopción de una postura decidida conocida popularmente como «pose de poder» puede inducir cambios hormonales y de otro tipo—.[11] El uso inapropiado de métodos estadísticos estándar ha recibido una buena parte de las críticas de la que es conocida como la crisis de reproductibilidad o replicación en la ciencia.

Con la creciente disponibilidad de bases de datos masivas y de programas informáticos de fácil uso, se podría pensar que hay una menor necesidad de formación en métodos estadísticos. Pensar eso sería, sin embargo, muy ingenuo. Lejos de liberarnos de la necesidad de adquirir conocimientos estadísticos, la disponibilidad de un mayor número de datos y el aumento del número y complejidad de los estudios científicos dificulta aún más el alcanzar las conclusiones correctas. Más datos significa que necesitamos ser aún más conscientes del valor de la evidencia disponible.

Por ejemplo, unos análisis de bases de datos derivadas de fuentes rutinarias pueden incrementar la posibilidad de falsos descubrimientos, tanto por los sesgos sistemáticos inherentes a las fuentes de los datos como por el hecho de presentar, tras muchos análisis, solo los resultados que parecen más interesantes, una práctica que se conoce normalmente como «dragado de datos». Para poder criticar el trabajo científico publicado, y aún más los informes de los medios de comunicación que nos encontramos todos los días, deberíamos ser capaces de tener una consciencia aguda de los peligros de la presentación selectiva de resultados, la necesidad de que los descubrimientos científicos sean replicados por parte de investigadores independientes, y el peligro de sobreinterpretar un único estudio fuera de contexto.

Todas estas ideas pueden englobarse bajo el término alfabetización en datos, que describe la capacidad no solo de llevar a cabo análisis estadísticos de problemas del mundo real, sino también de comprender y criticar cualquier conclusión alcanzada por otros basada en análisis estadísticos. Pero mejorar esa alfabetización en datos exige cambiar la forma en la que se enseña la estadística.

La enseñanza de la estadística

Generaciones de estudiantes han padecido áridos cursos de estadística basados en el aprendizaje de un conjunto de técnicas aplicadas a distintas situaciones, centrados más en la teoría matemática que en la compresión de por qué se usan unas determinadas fórmulas y en los problemas que surgen cuando se intentan usar datos para responder a

preguntas.

Afortunadamente, esto está cambiando. Las necesidades de la ciencia de los datos y de la alfabetización en datos demandan un enfoque más dirigido a la resolución de problemas, en el cual la aplicación de unas determinadas herramientas estadísticas sea vista como un componente más de un ciclo completo de investigación. En el libro adoptaremos la estructura PPDAC[12] como una forma de representar un ciclo de resolución de problemas.[13] La figura 0.3 se basa en un ejemplo de Nueva Zelanda, que es líder mundial en la enseñanza de la estadística en las escuelas.

El primer estadio del ciclo es especificar un Problema; la investigación estadística siempre empieza con una pregunta, como cuando nos preguntábamos por la pauta de los asesinatos cometidos por Harold Shipman o por el número de árboles que hay en el mundo. Más adelante nos centraremos en problemas que van desde el beneficio esperado de distintas terapias posteriores a una cirugía de cáncer de mama, a por qué los hombres mayores tienen las orejas grandes.

Es tentador saltarse la elaboración de un Plan detallado. La pregunta relacionada con el caso Shipman simplemente requería la reunión del mayor número de datos posible sobre sus víctimas. Pero los investigadores que se dedicaron a contar árboles dedicaron mucha atención a establecer unas definiciones precisas y a cómo llevar a cabo sus mediciones, dado que solo se pueden extraer conclusiones creíbles de un estudio que ha sido diseñado correctamente. Desgraciadamente, cuando hay prisa por encontrar datos adecuados y comenzar el análisis, a menudo se pasa por alto el diseño de la investigación.

La recopilación de buenos Datos requiere el tipo de conocimientos organizativos y de codificación que cada vez se valoran más en la ciencia de datos, especialmente debido a que los datos provenientes de fuentes cotidianas normalmente necesitan mucha limpieza antes de poder analizarse. Los sistemas de recopilación de datos pueden haber cambiado con el tiempo, podría haber errores obvios y así sucesivamente —la frase «datos encontrados» ya apunta a que pueden ser datos bastante desordenados, como algo que hemos cogido de la calle—.

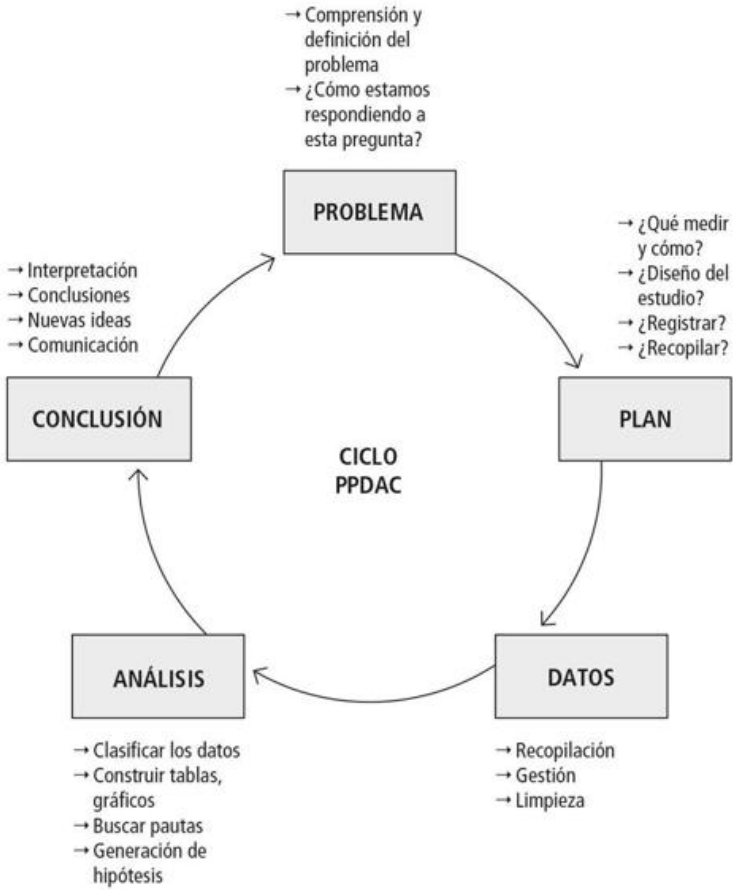


Figura 0.3. Ciclo PPDAC de resolución de problemas, que va del Problema, el Plan, los Datos, el Análisis, a las Conclusiones y la comunicación, y vuelta a empezar en un nuevo ciclo.

La fase de Análisis ha sido la que tradicionalmente ha recibido mayor atención en los cursos de estadística, y, de hecho, en este libro nos ocuparemos de una amplia variedad de técnicas analíticas. Pero en ocasiones, basta con una visualización bien presentada, como en el caso de la figura 0.1. Finalmente, la clave de la buena ciencia estadística es la extracción de unas Conclusiones adecuadas que reconozcan plenamente las limitaciones en la evidencia, y que sean comunicadas claramente, como en las ilustraciones gráficas de los datos de Shipman. Cualquier conclusión normalmente plantea más preguntas, y, de esta manera, el ciclo vuelve a empezar, como cuando empezamos a investigar la hora del día en la que murieron los pacientes de Shipman.

Aunque en la práctica el ciclo PPDAC de la figura 0.3 podría no seguirse al pie de la letra, lo que refleja es que las técnicas formales de análisis estadístico solo juegan una parte en el trabajo de un estadístico o de un científico cuantitativo. La ciencia estadística es mucho más que una rama de las matemáticas plagada de fórmulas esotéricas a las que se han enfrentado (a menudo a regañadientes) generaciones de estudiantes.

Este libro

Cuando era estudiante en Gran Bretaña en la década de 1970, solo había tres canales de televisión, los ordenadores eran del tamaño de un armario doble y lo más cercano que teníamos a Wikipedia era el aparato manual imaginado por Douglas Adams en su (sorprendentemente profética) novela Guía del autoestopista galáctico. Por ello, para mejorar nuestros conocimientos recurríamos a los libros de la colección Pelican, y sus icónicos lomos azules eran omnipresentes en las estanterías de todo estudiante.

Como estaba estudiando Estadística, mi colección de Pelican incluía Facts from Figures (Hechos derivados de las cifras), de M. J. Moroney (1951), y Cómo mentir con estadísticas, de Darrell Huff (1954). Estas venerables publicaciones se vendían por cientos de miles, lo que reflejaba tanto el grado de interés por la estadística como la deprimente falta de opciones que había en esa época. Lo cierto es que estos clásicos han resistido notablemente bien el paso de los sesenta y cinco años transcurridos, pero los tiempos actuales demandan un enfoque diferente a la enseñanza de la estadística, basada en los principios planteados más arriba.

Este libro, por tanto, utiliza la resolución de problemas del mundo real como punto de partida para la enseñanza de las ideas estadísticas. Algunas de estas ideas podrían parecer obvias, pero otras son más sutiles y requieren cierto esfuerzo mental, aunque no se necesitan conocimientos matemáticos. Comparado con los textos tradicionales, este libro se centra en cuestiones conceptuales más que en tecnicismos, e incluye solo unas pocas ecuaciones inofensivas, que cuentan con el apoyo de un glosario. Los programas informáticos son una parte vital de cualquier trabajo en el campo de la ciencia de los datos y de la estadística, pero no son el objeto central de este libro —existen tutoriales fácilmente accesibles para programas gratuitos como R y Python—.

Las preguntas incluidas en los recuadros pueden, hasta cierto punto, ser respondidas mediante análisis estadísticos, aunque su alcance varía

mucho. Algunas son importantes hipótesis científicas, como la referida a la existencia del bosón de Higgs, o a si hay evidencia convincente sobre la percepción extrasensorial (PES). Otras son preguntas sobre la asistencia sanitaria, como por ejemplo si los hospitales más ajetreados tienen unas tasas de supervivencia mayores, y si los chequeos para el cáncer de ovarios son beneficiosos. En ocasiones solo queremos estimar cantidades, como el riesgo de cáncer asociado a los bocadillos de beicon, el número de parejas sexuales que las personas que viven en Gran Bretaña han tenido a lo largo de su vida o el beneficio de tomar estatina diariamente.

Y algunas preguntas son simplemente interesantes, como identificar al superviviente más afortunado del Titanic, determinar si Harold Shipman podía haber sido detenido antes o estimar la probabilidad de que un esqueleto encontrado en un aparcamiento de Leicester sea realmente Ricardo III.

Este libro está dirigido tanto a estudiantes de Estadística que buscan una introducción no técnica a las cuestiones básicas, como a lectores en general que quieren estar más informados acerca de las estadísticas que se encuentran en su trabajo y en su vida diaria. Pondré el énfasis en el manejo adecuado y cuidadoso de la estadística: los números pueden parecer fríos y duros, pero los intentos mencionados de medir árboles, felicidad y fallecimientos ya nos han mostrado que tienen que tratarse con cuidado.

La estadística puede proporcionar claridad y perspectiva sobre los problemas a los que nos enfrentamos, pero todos estamos familiarizados con casos en los que se emplea de manera abusiva, a menudo para promover una determinada opinión o simplemente para llamar la atención. La capacidad de evaluar la fiabilidad de afirmaciones basadas en la estadística es un conocimiento clave en el mundo moderno, y espero que este libro sirva para conferir a la gente la capacidad de cuestionar los números con los que se encuentran en su vida cotidiana.

Resumen

→ Convertir experiencias en datos no es algo sencillo, y los datos tienen, inevitablemente, una capacidad limitada para describir el mundo.

→ La ciencia estadística tiene una historia larga y exitosa, pero ahora está cambiando a la luz de la creciente disponibilidad de datos.

→ Tener conocimientos en estadística es una parte importante de lo que significa ser un científico de datos.

→ La enseñanza de la estadística está cambiando de un enfoque centrado en los métodos matemáticos a otro basado en un ciclo completo de resolución de problemas.

→ El ciclo PPDAC proporciona un marco adecuado: Problema-Plan-Datos-Análisis-Conclusión y comunicación.

→ Estar alfabetizado en datos es crucial en el mundo moderno.

[1] La señal y el ruido, de Nate Silver (Penguin, 2012 [trad. cast. en Península, 2014]), es una excelente introducción a cómo puede aplicarse la ciencia estadística a la formulación de predicciones en los deportes y en otros ámbitos.

[2] Los datos del caso Shipman se analizan en mayor detalle en D. Spiegelhalter y N. Best, «Shipman's Statistical Legacy», Significance 1:1 (2004), pp. 10-12. Todos los documentos de la Investigación Pública están disponibles en https://webarchive.nationalarchives.gov.uk/ukgwa/20090808155110/http://www.the-shipman-inquiry.org.uk/reports.asp.

[3] Los términos en negrita aparecen en el glosario al final del libro, que proporciona definiciones básicas y técnicas.

[4] Spoiler: casi con total seguridad lo habría hecho.

[5] Esta cifra tiene un margen de error de 0,1 billones, lo que significa que los investigadores confían que la cifra real esté entre 2,94 y 3,14 billones (admito que me parece demasiado preciso, teniendo en cuenta los muchos supuestos adoptados en el modelo). También estiman que se talan cada año 15.000 millones (15.000.000.000) de árboles, y que el planeta ha perdido el 46 % de sus árboles desde el comienzo de la civilización humana. T. W. Crowther et al., «Mapping Tree Density at a Global Scale», Nature 525 (2015), pp. 201-205.

[6] E. J. Evans, Thatcher and Thatcherism, Routledge, 2013, p. 30.

[7] Changes to National Accounts: Inclusion of Illegal Drugs and Prostitution in UK National Accounts (Internet), Oficina de Estadísticas Nacionales, 2014.

[8] La Oficina de Estadísticas Nacionales del Reino Unido presenta

diversas medidas de bienestar en https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing.

[9] Lo cual, si yo estuviese en la media, me daría un motivo de esperanza.

[10] N. T. Nikas, D. C. Bordlee y M. Moreira, «Determination of Death and the Dead Doner Rule: A Survey of the Current Law on Brain Death», Journal of Medicine and Philosophy 41:3 (2016), pp. 237-256.

[11] J. P. Simmons y U. Simonsohn, «Power Posing: P-Curving the Evidence», Psychological Science 28 (2017), pp. 687-693. Para una refutación, véase A. J. C. Cuddy, S. J. Schultz y N. E. Fosse, «P-Curving a More Comprehensive Body of Research on Postural Feedback Reveals Clear Evidential Value for Power-Posing Effects: Reply to Simmons and Simonsohn (2017)», Psychological Science (2018), pp. 656-666.

[12] Por sus siglas en inglés, Problem, Plan, Data, Analysis, Conclusion. (N. del T.).

[13] La primera recomendación de la Asociación Americana de Estadística es «enseñar la estadística como un proceso de investigación para la resolución de problemas y la toma de decisiones». Véase https://www.amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx. El ciclo PPDAC fue desarrollado en R. J. MacKay y R. W. Oldford, «Scientific Method, Statistical Method and the Speed of Light», Statistical Science 15 (2000), pp. 254-278. Es promovido intensamente en el sistema educativo de Nueva Zelanda, que proporciona una muy buena formación en estadística. Véase C. J. Wild y M. Pfannkuch, «Statistical Thinking in Empirical Enquiry», International Statistical Review 67 (1999), pp. 223-265, y el curso online Data to Insight https://www.futurelearn.com/courses/data-to-insight.

01

Cómo entender las cosas como proporciones: datos categóricos y porcentajes

¿Qué les pasó a los niños que fueron operados del corazón en Bristol entre 1984 y 1995?

Joshua L. tenía 16 meses y padecía de transposición de las grandes arterias, una forma severa de enfermedad cardiaca congénita en la cual las principales arterias coronarias salen del ventrículo que no les corresponde. Necesitaba una operación para «intercambiar» las arterias. A las siete de la mañana del 12 de enero de 1995, sus padres se despidieron de él y lo vieron alejarse camino del quirófano en el Hospital Real de Bristol. Pero los padres de Joshua no tenían ni idea de las historias que circulaban desde comienzos de la década de 1990 acerca de las bajas tasas de supervivencia en ese hospital. Nadie les contó que muchas enfermeras habían dejado esa unidad para no tener que comunicar una y otra vez a los padres que su hijo había fallecido, o que la noche anterior había habido una reunión de última hora para decidir si había que cancelar la operación de Joshua.[14]

Joshua murió en el quirófano. Al año siguiente, el Consejo General Médico (el órgano regulador de la medicina) puso en marcha una investigación para responder a las denuncias de los padres de Joshua y de otros afligidos progenitores, y en 1998 dos cirujanos y el exdirector general fueron declarados culpables de mala práctica médica grave.

La opinión pública siguió con preocupación estos casos, por lo que se ordenó la puesta en marcha de una investigación oficial. Se formó un equipo de estadísticos para llevar a cabo la triste tarea de comparar las tasas de supervivencia en Bristol con las del resto del Reino Unido entre 1984 y 1995. Yo dirigía ese equipo.

En primer lugar, tuvimos que determinar cuántos niños habían sido sometidos a cirugía cardiaca, y cuántos de ellos habían fallecido. A primera vista, esto parece bastante sencillo, pero, como vimos en el capítulo anterior, algo tan aparentemente fácil como contar eventos puede llegar a ser complicado. ¿Qué es un «niño»? ¿Qué procedimiento cuenta como «cirugía cardiaca»? ¿Cuándo puede la muerte atribuirse a la cirugía? E incluso cuando se ha tomado una decisión acerca de estas definiciones, ¿podríamos determinar cuántas de cada ha habido?

Consideramos «niño» a alguien menor de 16 años, y nos centramos en la cirugía «a corazón abierto», en la cual se para el corazón y su función es reemplazada por un baipás cardiopulmonar. Podía haber múltiples operaciones por admisión, pero todas ellas eran consideradas como un solo evento. Las muertes se contaban si se producían en los treinta días posteriores a la operación, fuesen o no en el hospital o debidas a la cirugía. Sabíamos que la muerte era una medida imperfecta de la calidad del resultado, dado que dejaba fuera a niños que habían sufrido daño cerebral u otra discapacidad como resultado de la cirugía, pero no teníamos datos sobre resultados a largo plazo.

La principal fuente de datos era la Estadística de Episodios Hospitalarios (EEH), que se elaboraba a partir de datos administrativos introducidos por codificadores mal pagados. La EEH tenía mala reputación entre los médicos, pero era una fuente que tenía la ventaja de que podía cruzarse con los datos nacionales sobre fallecimientos. También existía un sistema paralelo de datos sometido directamente al Registro de Cirugía Cardiaca (RCC) establecido por la asociación profesional de cirujanos.

Estas dos fuentes de datos, aunque se suponía que se referían a lo mismo, mostraban considerables diferencias: para 1991-1995, la EEH incluía 62 muertes en 505 operaciones a corazón abierto (14 %), mientras que el RCC afirmaba que se habían producido 71 muertes en 653 operaciones (13 %). Estaban disponibles nada menos que otras cinco fuentes de datos locales, desde el historial de anestesias hasta los registros personales de los propios cirujanos. Bristol estaba repleto de datos, pero ninguna de las fuentes de datos podía considerarse «la verdad», y nadie asumía la responsabilidad de analizar y actuar en base a los resultados quirúrgicos.

Calculamos que, si los pacientes de Bristol tuviesen el riesgo medio prevaleciente en el resto del Reino Unido, se podrían haber esperado 32 muertes durante este período, en lugar de las 62 consignadas en la EEH, algo que incluimos como «30 muertes de más» entre 1991 y 1995.[15] El número exacto varía según la fuente de datos utilizada. Era increíble que no pudiésemos establecer ni siquiera los datos básicos sobre el número de operaciones y su resultado, aunque hay que reconocer que los actuales sistemas de recopilación de datos deben haber mejorado.

Estos descubrimientos tuvieron una amplia repercusión en la prensa, y la investigación sobre el hospital de Bristol provocó un gran cambio en las actitudes sobre la supervisión del desempeño clínico: a partir de

entonces, se dejó de confiar en que la profesión médica se controlase a sí misma. Se establecieron mecanismos para garantizar la publicidad de los datos de supervivencia hospitalaria, aunque, como veremos, la forma en la que se presentan estos datos puede influir en la percepción de los mismos.

Comunicar recuentos y proporciones

Los datos referidos a si eventos individuales han sucedido o no se denominan datos binarios, dado que solo tienen dos valores, generalmente etiquetados como sí y no. Los conjuntos de datos binarios pueden ser resumidos por el número de veces y el porcentaje de casos en los cuales se ha producido un evento.

La idea de este capítulo es que la presentación básica de los datos estadísticos es importante. En cierto sentido, estamos dando un salto hasta el último paso del ciclo PPDAC, en el cual se comunican las conclusiones, y aunque la forma que adquiere esta comunicación no ha sido considerada tradicionalmente un tema importante en estadística, el aumento del interés por la visualización de los datos refleja un cambio en esta actitud. En este capítulo y en el siguiente nos concentraremos en cómo presentar los datos para que se pueda captar la esencia de lo que está pasando sin necesidad de un análisis detallado, comenzando con un vistazo a las distintas formas de presentar unos datos que, en gran medida gracias a la investigación de Bristol, ahora están disponibles para el público.

Hospital	Número de bebés sometidos a cirugía	Número de supervivientes al menos 30 días después de la cirugía	Número de fallecidos durante los 30 días posteriores a la cirugía	Porcentaje de supervivientes	Porcentaje de fallecimientos
Londres, Harley Street	418	413	5	98,8	1,2
Leicester	607	593	14	97,7	2,3
Newcastle	668	653	15	97,8	2,2
Glasgow	760	733	27	96,3	3,7
Southampton	829	815	14	98,3	1,7
Bristol	835	821	14	98,3	1,7
Dublín	983	960	23	97,7	2,3
Leeds	1.038	1.016	22	97,9	2,1
Londres, Brompton	1.094	1.075	19	98,3	1,7
Liverpool	1.132	1.112	20	98,2	1,8
Londres, Evelina	1.220	1.185	35	97,1	2,9
Birmingham	1.457	1.421	36	97,5	2,5
Londres, Great Ormond Street	1.892	1.873	19	99,0	1,0
Total	12.933	12.670	263	98,0	2,0

Tabla 1.1. Resultados de la cirugía cardiaca infantil en hospitales del Reino Unido e Irlanda entre 2012 y 2015, en términos de supervivencia o no, a los 30 días posteriores a la operación.

La tabla 1.1 muestra los resultados para casi trece mil niños que fueron sometidos a cirugía cardiaca en el Reino Unido e Irlanda entre 2012 y 2015.[16] Doscientos sesenta y tres bebés murieron en los treinta días posteriores a ser operados, y cada una de estas muertes es una tragedia para sus familias. No será mucho consuelo para ellos saber que las tasas de supervivencia han mejorado enormemente desde la época de la investigación de Bristol, que ahora de media son un 98 %, y que, por tanto, las perspectivas son mucho mejores para las familias de los niños que se enfrentan a una cirugía cardiaca.

Una tabla puede ser considerada como un tipo de gráfico, y requiere un diseño cuidadoso del color, la fuente y el lenguaje para asegurarnos de que sea atractiva y legible. La respuesta emocional de la audiencia a la tabla puede estar influida también por la elección de qué columnas presentar.

La tabla 1.1 muestra los resultados en términos de supervivientes y fallecimientos, pero mientras que en Estados Unidos se proporcionan datos de las tasas de mortalidad de la cirugía cardiaca infantil, en el Reino Unido se proporcionan las tasas de supervivencia. Esto se conoce como marcos negativos o positivos. Su efecto global en nuestra percepción es intuitivo, y está muy documentado: «un 5 % de mortalidad» suena peor que un «95 % de supervivencia». Proporcionar el número de fallecimientos además del porcentaje puede también incrementar la percepción de riesgo, al imaginarnos ese total como un montón de personas reales.

Un ejemplo clásico de cómo unos marcos alternativos pueden cambiar el impacto emocional de una cifra es un anuncio que apareció en el metro de Londres en 2011, que afirmaba que «el 99 % de los jóvenes londinenses no cometen violencia juvenil grave». La intención de estos anuncios era presumiblemente hacer sentir a los pasajeros que vivían en una ciudad segura, pero se podía dar la vuelta a su impacto emocional con solo dos sencillos cambios. En primer lugar, esa afirmación suponía que el 1 % de los jóvenes londinenses sí cometen violencia grave. En segundo lugar, dado que la población de Londres es de alrededor de 9 millones, hay aproximadamente 1 millón de personas con edades comprendidas entre los 15 y los 25 años, y si consideramos a estas personas como «jóvenes», esto significa que hay un 1 % de 1 millón, es decir, 10.000 jóvenes muy violentos en la ciudad. Esto, la verdad, no es muy tranquilizador. Nótense los dos trucos utilizados para manipular el impacto de esta estadística: pasar de un marco positivo a uno negativo, y después pasar de un porcentaje al número real de personas.

Idealmente, deberían presentarse tanto los marcos positivos como los negativos, si queremos proporcionar una información imparcial, aunque el orden de las columnas podría, aun así, influir en la interpretación de la tabla. El orden de las filas de una tabla también tiene que ser considerado detenidamente. La tabla 1.1 muestra los hospitales ordenados según el número de operaciones en cada uno de ellos, pero si se mostrasen, por ejemplo, ordenados por las tasas de mortalidad, con aquellos con las más altas en la parte superior de la tabla, podríamos dar la impresión de que se trata de una forma válida y relevante de comparar hospitales. Esas tablas clasificatorias son las favoritas de los medios y de algunos políticos, pero pueden ser muy engañosas: no solo porque las diferencias pueden deberse a variación aleatoria, sino también porque puede que los hospitales estén tratando casos muy diferentes. En la tabla 1.1, por ejemplo, podríamos sospechar que Birmingham, uno de los mayores y más famosos hospitales infantiles, es el que se ocupa de los casos más graves, y por

ello sería injusto, por decirlo suavemente, destacar sus aparentemente pobres tasas de supervivencia.[17]

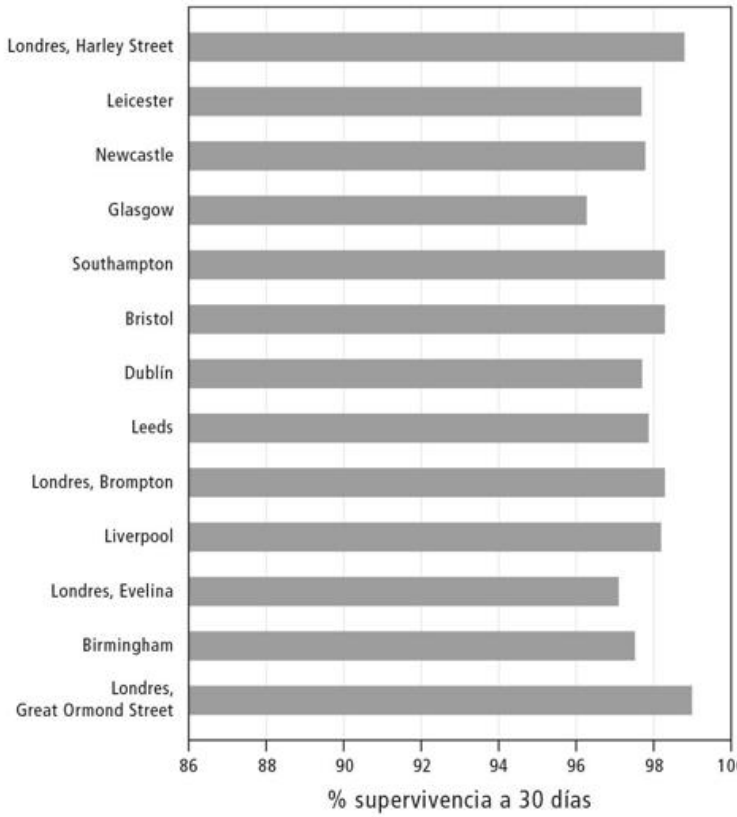


Figura 1.1. Diagrama de barras horizontales de las tasas de supervivencia a 30 días para 13 hospitales. La elección del comienzo del eje horizontal, que en este caso es el 86 %, puede tener un efecto crucial sobre la impresión dada por el gráfico. Si el eje comienza en el 0 %, todos los hospitales parecerán indistinguibles, mientras que si empezamos en el 95 % las diferencias parecerían engañosamente enormes. En lugar de un diagrama de barras, sería mejor usar puntos para cada dato cuando el eje no empieza en el cero.

Las tasas de supervivencia pueden ser presentadas en un diagrama de barras horizontales como el que se muestra en la figura 1.1. Una decisión de suma importancia es dónde comenzar el eje horizontal: si los valores empiezan en el 0 %, todas las barras tendrán una longitud similar a la del gráfico, algo que claramente mostrará las extraordinariamente altas tasas de supervivencia, pero las líneas serán indistinguibles. Pero el truco más viejo para engañar con un gráfico

sería empezar el eje, por ejemplo, en el 95 %, algo que haría que los hospitales pareciesen ser extremadamente distintos, aunque, de hecho, solo hubiese una pequeña variación aleatoria.

Por tanto, decidir dónde comienza el eje representa un dilema. Alberto Cairo, autor de libros influyentes sobre visualización de datos,[18] sugiere que siempre hay que empezar con un «valor de base lógico y significativo», que en este caso parece difícil de identificar —mi decisión bastante arbitraria del 86 % representa aproximadamente la inaceptablemente baja tasa de supervivencia de Bristol hace veinte años—.

He comenzado este libro con una cita de Nate Silver, el fundador de la plataforma de datos FiveThirtyEight, famoso por haber predicho acertadamente los resultados de las elecciones presidenciales estadounidenses de 2008, y que expresó elocuentemente la idea de que los números no hablan por sí mismos: somos nosotros los que debemos darles significado. Esto implica que la comunicación es una parte clave del ciclo de resolución de problemas, y he mostrado en esta sección cómo el mensaje que nos comunica un conjunto sencillo de porcentajes puede verse influido por nuestras decisiones sobre cómo presentarlo.

Ahora debemos introducir un concepto práctico e importante que nos ayudará a ir un paso más allá de las sencillas preguntas de sí o no.

Variables categóricas

Una variable es una medida que puede adoptar diferentes valores en diferentes circunstancias; es un término muy conveniente para referirse a todos los tipos de observaciones que comprenden datos. Las variables binarias son cuestiones de sí o no, como por ejemplo si alguien está vivo o muerto, o si es una mujer o no: ambas cuestiones varían entre personas, y pueden, incluso en el caso del género, variar a lo largo del tiempo para la misma persona. Las variables categóricas son medidas que pueden adoptar dos o más categorías, que pueden ser, a su vez:

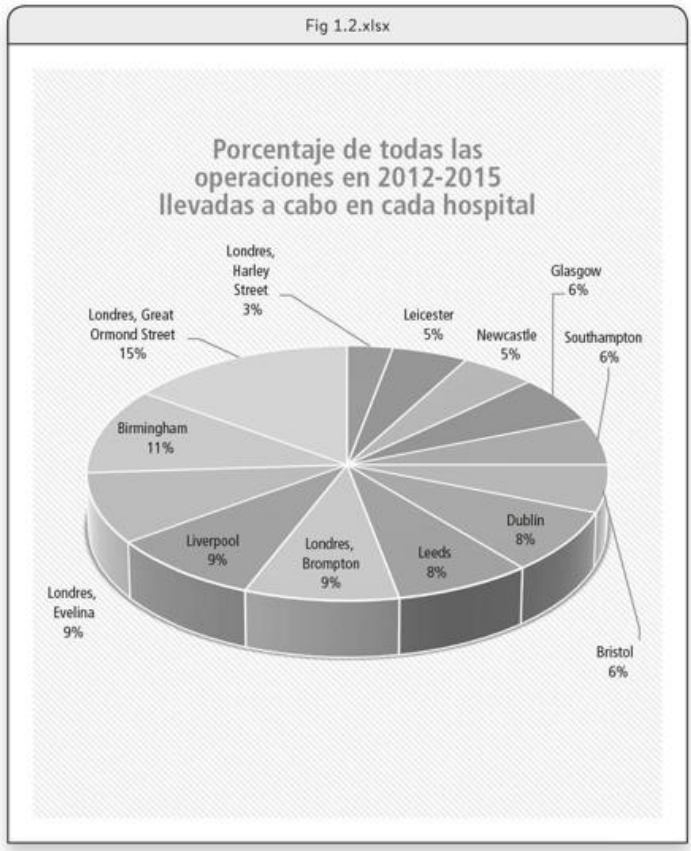


Figura 1.2. Proporción de todas las operaciones infantiles de corazón llevadas a cabo en cada hospital, desplegadas en un gráfico de tarta de Excel. Este gráfico profundamente desagradable hace que las categorías que aparecen al frente parezcan más grandes, lo que hace imposible realizar una comparación visual entre hospitales.

→ Categorías no ordinales: como, por ejemplo, el país de origen de una persona, el color de un coche o el hospital en el que tiene lugar una operación.

→ Categorías ordinales: como, por ejemplo, los rangos del personal militar.

→ Números que han sido agrupados: como los niveles de obesidad, que se definen normalmente en términos de umbrales del índice de masa corporal (IMC).[19]

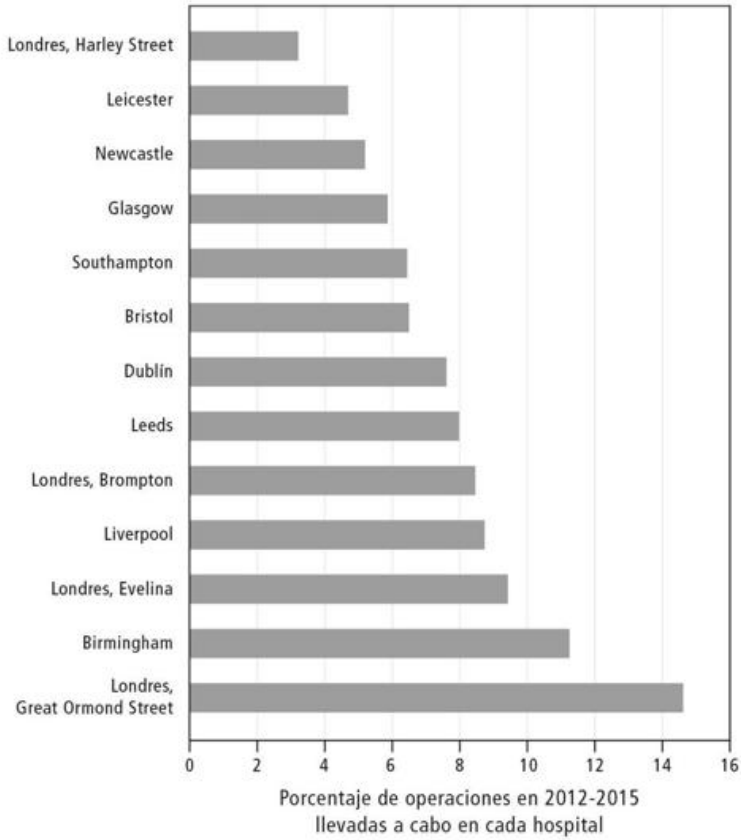


Figura 1.3. Porcentaje de todas las operaciones infantiles de corazón llevadas a cabo en cada hospital: una representación más clara mediante un diagrama de barras horizontales.

Para la presentación de datos categóricos, los gráficos de tarta permiten hacerse una impresión del tamaño de cada categoría en relación con la totalidad de la tarta, pero muchas veces son visualmente confusos, especialmente si intentan mostrar demasiadas categorías en el mismo gráfico, o usan una representación tridimensional que distorsiona las áreas. La figura 1.2 es un ejemplo especialmente lamentable del tipo de gráfico de tarta ofrecido por Microsoft Excel. Muestra las proporciones de los 12.933 pacientes infantiles de corazón de la tabla 1.1 tratados en cada hospital.

Utilizar múltiples gráficos de tarta normalmente no es una buena idea, porque las comparaciones se ven perjudicadas por la dificultad en evaluar los tamaños relativos de áreas de formas diversas. Es mejor basar las comparaciones solamente en la altura o la longitud de un diagrama de barras. La figura 1.3 muestra un ejemplo más sencillo y

más claro de un diagrama de barras horizontales del porcentaje tratado en cada hospital.

Comparación de un par de proporciones

Hemos visto cómo un conjunto de proporciones puede ser elegantemente comparado utilizando un diagrama de barras, por lo que sería razonable pensar que comparar dos proporciones sería algo trivial. Pero cuando estas proporciones representan estimadores de los riesgos de experimentar un determinado daño, entonces la forma en la cual esos riesgos son comparados se convierte en una cuestión seria y polémica. Veamos una pregunta típica en este sentido:

¿Cuál es el riesgo de cáncer derivado de comer bocadillos de beicon?

Todos conocemos los titulares sensacionalistas de los medios que advierten de que algo de lo más cotidiano incrementa el riesgo de que algo malo nos ocurra; a mí me gusta llamarlas historias del tipo «los gatos provocan cáncer». Por ejemplo, en noviembre de 2015 el Centro Internacional de Investigaciones sobre el Cáncer de la Organización Mundial de la Salud anunció que la carne procesada constituía un «carcinógeno de grupo I», lo que la situaba en la misma categoría que los cigarrillos y el amianto. Esto, inevitablemente, generó titulares histéricos como el del Daily Record, que afirmaba: «Los expertos advierten que el beicon, el jamón y las salchichas tienen el mismo riesgo de provocar cáncer que los cigarrillos».[20]

La IARC[21] intentó detener el escándalo señalando que la clasificación como grupo 1 suponía simplemente que hay un aumento del riesgo de cáncer, pero no dice nada sobre la magnitud de ese riesgo. En la letra pequeña de las noticias de prensa se recogía que, según la IARC, el consumo de 50 gramos de carne procesada al día estaba asociado con un aumento del 18 % del riesgo de cáncer colorrectal. Esto suena preocupante, pero ¿lo es realmente?

La cifra del 18 % se conoce como riesgo relativo dado que representa el incremento del riesgo de tener cáncer colorrectal entre un grupo de personas que consumen 50 gramos de carne procesada al día, lo que supone, por ejemplo, un bocadillo diario de dos lonchas de beicon, y un grupo que no. Los expertos en estadística toman este riesgo relativo y lo reinterpretan como un cambio en el riesgo absoluto, lo que significa un cambio en el porcentaje real de cada grupo que se esperaría que sufriese un evento adverso.

Su conclusión fue que, en condiciones normales, se podría esperar que

6 de cada 100 personas que no tomasen un bocadillo de beicon diario desarrollasen cáncer colorrectal a lo largo de su vida. Si 100 personas similares comiesen un bocadillo de beicon cada día de su vida, entonces, de acuerdo con el informe de la IARC, esperaríamos que un 18 % más desarrollase un cáncer colorrectal, lo que supondría pasar de 6 a 7 casos de cada 100.[22] Esto supone un caso adicional de cáncer colorrectal entre esos 100 consumidores diarios de beicon, lo que no parece tan impresionante como el riesgo relativo (un incremento del 18 %), y podría servir para poner este riesgo en perspectiva. Necesitamos distinguir lo realmente peligroso de lo que suena aterrador.[23]

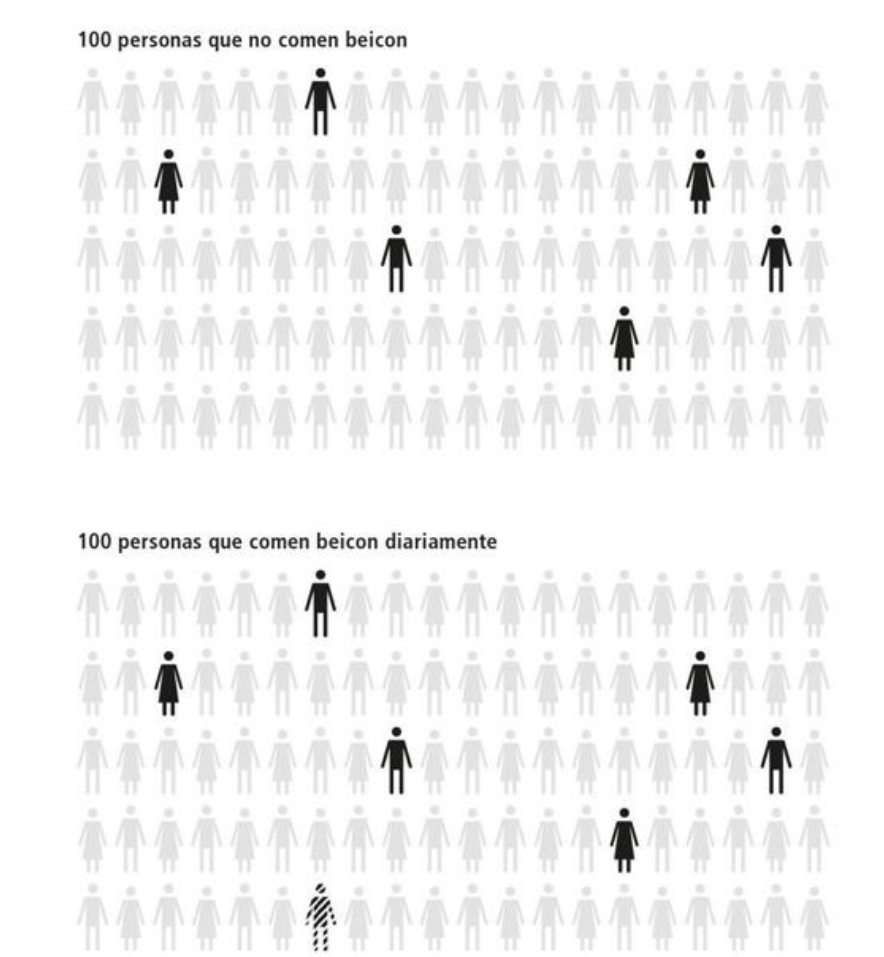


Figura 1.4. El ejemplo del bocadillo de beicon utilizando un par de matrices de iconos, con iconos dispersos aleatoriamente y mostrando el incremento del riesgo de cáncer por el consumo diario de beicon. De 100 personas que no comen beicon, 6 (iconos en negro)

desarrollan cáncer colorrectal bajo condiciones normales. De 100 personas que comen beicon diariamente, hay un caso adicional (el icono a rayas). (En puridad, los 6 iconos en negro deberían estar situados de manera diferente en los dos gráficos, dado que representan 2 grupos distintos de 100 personas. Pero esto haría más difícil la comparación entre ambos grupos).

El ejemplo del bocadillo de beicon muestra las ventajas de comunicar el riesgo con frecuencias esperadas: en lugar de discutir porcentajes o probabilidades, simplemente nos preguntamos: «¿Qué significa esto para 100 (o 1.000) personas?». Los estudios psicológicos han mostrado que esta técnica mejora la comprensión; de hecho, comunicar que este consumo de carne adicional llevaba a un «aumento del riesgo de un 18 %» podría considerarse manipulador, dado que sabemos que esta forma de presentar la información da una impresión exagerada de la importancia del riesgo.[24] La figura 1.4 usa una matriz de iconos para representar directamente las frecuencias esperadas de cáncer colorrectal por cada 100 personas.

En la figura 1.4, los iconos del «cáncer» están dispersos aleatoriamente entre los 100. Aunque esta dispersión incrementa la impresión de impredecibilidad, solo debe ser empleada cuando hay un único icono destacado. No debería ser necesario contar iconos para hacer una rápida comparación visual.

La tabla 1.2 muestra otras formas de comparar dos porcentajes, de nuevo referidos al riesgo para las personas que comen beicon y las que no.

«1 de X» es una forma normal de expresar el riesgo, como por ejemplo si hablamos de «1 de cada 16 personas», para representar un riesgo del 6 %. Pero usar múltiples afirmaciones del tipo «1 de …» no es recomendable, porque para mucha gente son difíciles de comparar. Por ejemplo, a la pregunta: «¿Cuál es el mayor riesgo, 1 de 100, 1 de 10 o 1 de 1.000?», en torno a una cuarta parte de los interpelados responden erróneamente; el problema es que el número más alto se asocia con el riesgo más bajo, por lo que se requiere cierta pericia mental para no equivocarse.

Técnicamente, las probabilidades de un evento es la razón entre la probabilidad de que el evento suceda y la probabilidad de que no suceda. Por ejemplo, dado que de 100 personas que no consumen beicon, 6 desarrollarán cáncer colorrectal y 94 no, las probabilidades de desarrollar cáncer colorrectal en este grupo son 6/94, o, como se dice normalmente, «6 de 94». Las probabilidades se usan de forma

habitual, pero también se usan ampliamente en los modelos estadísticos de porcentajes, y esto significa que la investigación médica expresa normalmente los efectos asociados con tratamientos o comportamiento en términos de razones de probabilidad.

Método	No consumidores de beicon	Consumidores de beicon diarios
Tasa de evento	6 %	7 %
Frecuencia esperada	6 de 100	7 de 100
	1 de 16	1 de 14
Probabilidades	6/94	7/93

Medidas comparadas	
Diferencia absoluta de riesgo	1 %, o 1 de 100
Riesgo relativo	1,18, o un incremento del 18 %
«Número que es necesario tratar»	100
Razón de probabilidad	(7/93) / (6/94) = 1,18

Tabla 1.2. Ejemplos de métodos para comunicar el riesgo de cáncer colorrectal a lo largo de la vida con y sin consumo diario de beicon. El «número que es necesario tratar» es el número de personas que se necesita que coman un bocadillo de beicon todos los días de su vida para esperar un caso adicional de cáncer colorrectal (por lo que quizá podría llamarse mejor «número que es necesario que coman»).

Aunque son extremadamente comunes en la investigación, las razones de probabilidad son una forma bastante poco intuitiva de resumir las diferencias de riesgos. Si los eventos son muy raros, entonces la razón de probabilidad será numéricamente cercana a los riesgos relativos, como en el caso de los bocadillos de beicon, pero para eventos comunes la razón de probabilidad puede ser muy diferente del riesgo relativo, y el siguiente ejemplo muestra que esto puede ser muy confuso para los periodistas (y para otras personas).

¿Por qué pasar del 85 % al 87 % puede considerarse un incremento del 20 %?

Se asume universalmente que la estatina reduce el colesterol y el riesgo de ataques al corazón e ictus, pero algunos médicos se han mostrado preocupados por sus efectos secundarios. Un estudio publicado en 2013 descubrió que el 87 % de los consumidores de estatina sufrieron dolor muscular, frente a un 85 % de los que no tomaron estatina. Echando un vistazo a las distintas formas de comparar riesgos de la tabla 1.2, podríamos informar o bien de un

incremento de un 2 % en el riesgo absoluto, o de un riesgo relativo de 0,87 / 0,85 = 1,02, es decir, un incremento relativo del riesgo de un 2 %. Las probabilidades en los dos grupos son 0,87 / 0,13 = 6,7 y 0,85 *0,15 = 5,7, por lo que la razón de probabilidad será 6,7* 5,7 = = 1,18: exactamente la misma que la de los bocadillos de beicon, pero basada en riesgos absolutos muy distintos.

El Daily Mail malinterpretó esta razón de probabilidad de 1,18 como un riesgo relativo, y publicó un titular que afirmaba que la estatina «aumenta el riesgo hasta un 20 por ciento», que es una interpretación realmente errónea de lo que el estudio había descubierto. Pero no toda la culpa está del lado de los periodistas: el resumen del artículo mencionaba solamente la razón de probabilidad, sin mencionar que correspondía a una diferencia entre riesgos absolutos de 85 % versus 87 %.[25]

Esto pone de manifiesto el peligro de usar las razones de probabilidad fuera de un contexto científico, y la ventaja de presentar los riesgos absolutos siempre que uno se dirija a una audiencia, ya sean estos referidos al beicon, la estatina o a cualquier otra cosa.

Los ejemplos en este capítulo han demostrado cómo la tarea aparentemente sencilla de calcular y comunicar porcentajes puede ser algo complejo. Necesita llevarse a cabo con atención y cuidado; el impacto de los descriptivos de datos, numéricos o en gráficos, puede ser analizado con la ayuda de psicólogos con conocimientos en evaluación de la percepción de formatos alternativos. La comunicación es una parte importante del ciclo de resolución de problemas, y no debería ser simplemente una cuestión de preferencias personales.

Resumen

→ Las variables binarias son preguntas de sí o no, y pueden ser resumidas en porcentajes.

→ Los marcos positivos o negativos de los porcentajes pueden cambiar su impacto emocional.

→ Los riesgos relativos tienden a transmitir una importancia exagerada, por lo que, para mayor claridad, hay que proporcionar los riesgos absolutos.

→ Las frecuencias esperadas favorecen la comprensión y un sentido apropiado de la importancia de los datos.

→ La razón de probabilidad surge de los estudios científicos, pero no debe ser usada para la comunicación en general.

→ Hay que escoger con cuidado los gráficos y ser conscientes de su impacto.

[14] Véase «History of Scandal», Daily Telegraph, 18 de julio de 2001, y D. J. Spiegelhalter *et al.*, «Commissioned Analysis of Surgical Performance Using Routine Data: Lesson from the Bristol Inquiry», Journal of the Royal Statistical Society: Series A (Statistics in Society) 165 (2002), pp. 191-221.

[15] Ahora me arrepiento de haber usado el término «muertes de más», porque los periódicos lo interpretaron como «muertes evitables». En realidad, alrededor de la mitad de los hospitales tienen más muertes de las esperadas, por pura aleatoriedad, y solo unas pocas de esas muertes podrían ser consideradas evitables.

[16] Los datos sobre los resultados de la cirugía cardiaca en el Reino Unido pueden obtenerse en http://childrensheartsurgery.info/.

[17] No hay, en realidad, evidencia de diferencias sistemáticas entre estos hospitales, una vez que controlamos por la gravedad de sus casos.

[18] Véanse A. Cairo, The Truthful Art: Data, Charts, and Maps for Communication, New Riders, 2016, y The Functional Art: An Introduction to Information Graphics and Visualization, New Riders, 2012.

[19] El índice de masa corporal fue desarrollado por el estadístico belga Adolphe Quetelet antes de 1850, y se define como IMC = peso (kg) / estatura (m)2. El índice se desagrega en la actualidad en múltiples grupos. En el Reino Unido, las definiciones actuales de la obesidad son: peso inferior al normal (IMC < 18,5 kg/m2); normal (IMC entre 18,5 y 25); sobrepeso (entre 25 y 30); obesidad (de 30 a 35) y obesidad mórbida (por encima de 35).

[20] Organización Mundial de la Salud. Preguntas y respuestas acerca del riesgo de cáncer por consumo de carne roja y procesada en http://www.who.int/features/qa/cancer-red-meat/en/. «Los expertos advierten que el beicon, el jamón y las salchichas tienen el mismo riesgo de provocar cáncer que los cigarrillos», Daily Record, 23 de octubre de 2015.

[21] Por sus siglas en inglés, International Agency for Research in Cancer. (N. del T.).

[22] En puridad, un incremento del 18 % sobre un 6 % es 6 % × 1,18 = 7,08 %, pero redondearlo al 7 % es admisible para este nivel divulgativo.

[23] Este era el comentario favorito de Hans Rosling (véase el siguiente capítulo).

[24] E. A. Akl et al., «Using Alternative Statistical Formats for Presenting Risks and Risks Reductions», Cochrane Database of Systematic Review 3 (2011).

[25] «Statins Can Weaken Muscles and Joints: Cholesterol Drug Raises Risk of Problems by up to 20 per cent», Mail Online, 3 de junio de 2013. El estudio original es I. Mansi et al., «Statin and Musculoskeletal Conditions, Arthropathies, and Injuries», JAMA Internal Medicine 173 (2013), pp. 1318-1326.

02

Cómo resumir y comunicar los números.

Muchísimos números

¿Podemos confiar en la sabiduría de las multitudes?

En 1907, Francis Galton, primo de Charles Darwin y erudito pionero de la identificación mediante huellas dactilares, la previsión del tiempo y la eugenesia,[26] escribió una carta a la prestigiosa revista científica Nature sobre su visita a una exhibición de ganado voluminoso y aves de corral en la ciudad portuaria de Plymouth. Ahí había visto cómo la gente pagaba seis peniques para intentar adivinar el peso «eviscerado» de un gran buey ya descuartizado. Galton se hizo con 787 boletos y escogió el valor medio de 1.207 libras (547 kilos) como la decisión democrática, «cualquier otra estimación era rechazada como demasiado alta o demasiado baja por la mayoría de los votantes». El peso neto resultó ser de 1.198 libras (543 kilos), una cifra asombrosamente cercana a su estimación basada en los 787 votos.[27] Galton tituló su carta «Vox Populi» (voz del pueblo), pero este proceso de toma de decisiones es ahora mejor conocido como la sabiduría de las multitudes.

Galton llevó a cabo lo que llamaríamos un resumen de los datos: cogió un conjunto de números escritos en boletos y los redujo a un único peso estimado de 1.207 libras. En este capítulo echaremos un vistazo a las técnicas desarrolladas en el siglo siguiente para resumir y comunicar las montañas de datos que hay ahora disponibles. Veremos que los resúmenes numéricos de posición, dispersión, tendencia y correlación están íntimamente relacionados con la presentación de los datos en un gráfico tanto en papel como en una pantalla. Y nos ocuparemos de la suave transición de la simple descripción de los datos al relato de una historia a través de la infografía.

En primer lugar, empezaremos con mi propio intento de experimento de sabiduría de las multitudes, que muestra muchos de los problemas que surgen cuando se usa el indisciplinado mundo real, con toda su capacidad para la rareza y el error, como fuente de datos.

La estadística no solo se ocupa de eventos serios como el cáncer y la cirugía. El divulgador matemático James Grime y yo llevamos a cabo un experimento bastante trivial: subimos un vídeo a YouTube en el que se mostraba un tarro lleno de gominolas y preguntamos cuántas

había. Intente usted mismo adivinar el número de gominolas en la foto de la figura 2.1 (la cifra real se revelará más adelante). Novecientas quince personas nos dieron sus estimaciones, que iban de 219 a 31.337; en este capítulo veremos cómo pueden representarse gráficamente y resumirse numéricamente las variables de este tipo.

Para empezar, la figura 2.2 muestra tres formas de representar la pauta que siguen los valores proporcionados por los 915 respondientes. Estas pautas pueden ser denominadas distribución de datos, distribución muestral o distribución empírica.[28]

Figura 2.1. ¿Cuántas gominolas hay en este tarro? Hicimos esta pregunta en un vídeo de YouTube y obtuvimos 915 respuestas. La respuesta correcta se dará más adelante.

(a) El diagrama de cintas o diagrama de dispersión muestra simplemente cada dato específico como un punto, pero cada uno de ellos recibe una fluctuación de retardo para evitar que múltiples

estimaciones del mismo número se acumulen unas encima de otras oscureciendo la pauta general. Se muestra claramente un gran número de estimaciones con un rango de hasta 3.000, y después una larga «cola» de valores hasta alrededor de 30.000, con un agrupamiento exactamente en el valor 10.000.

(b) El diagrama de caja y bigotes resume algunas características esenciales de la distribución de los datos.[29]

(c) Este histograma simplemente nos cuenta cuántos datos específicos se encuentran en cada conjunto de intervalos —nos da una idea muy aproximada de la forma de la distribución—.

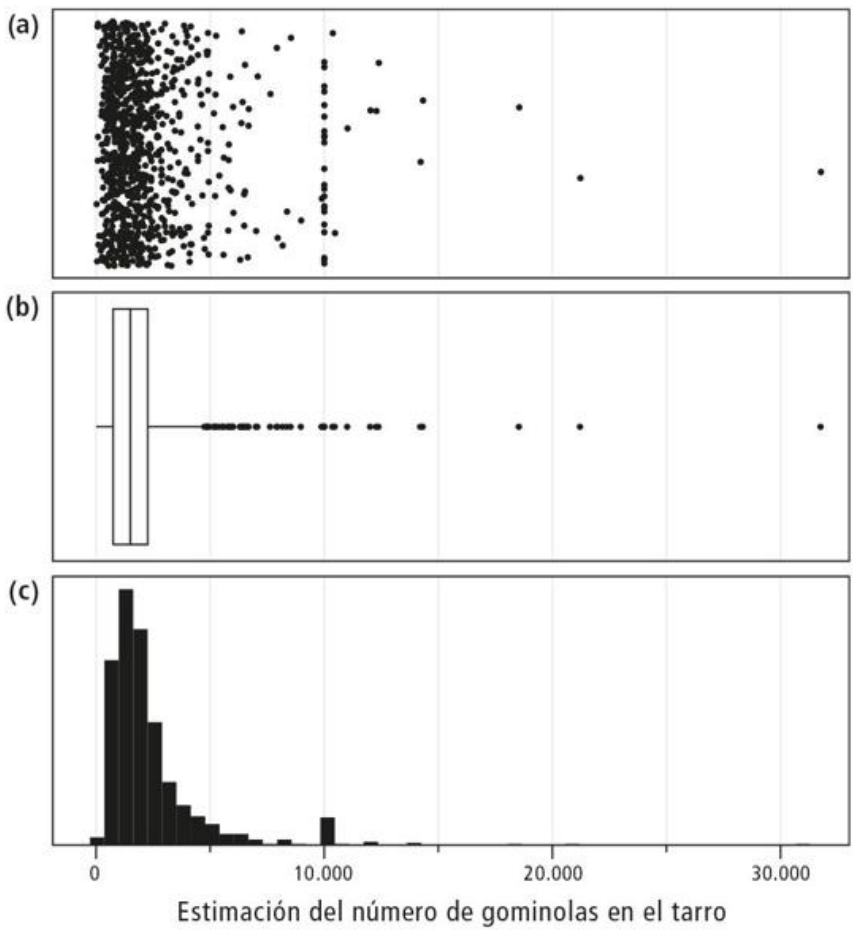


Figura 2.2. Diferentes formas de presentar la pauta de 915 estimaciones del número de gominolas en el tarro. (a) Diagrama de cintas o diagrama de dispersión con una fluctuación de retardo para

evitar que los puntos estén unos encima de otros; (b) diagrama de caja y bigotes; (c) histograma.

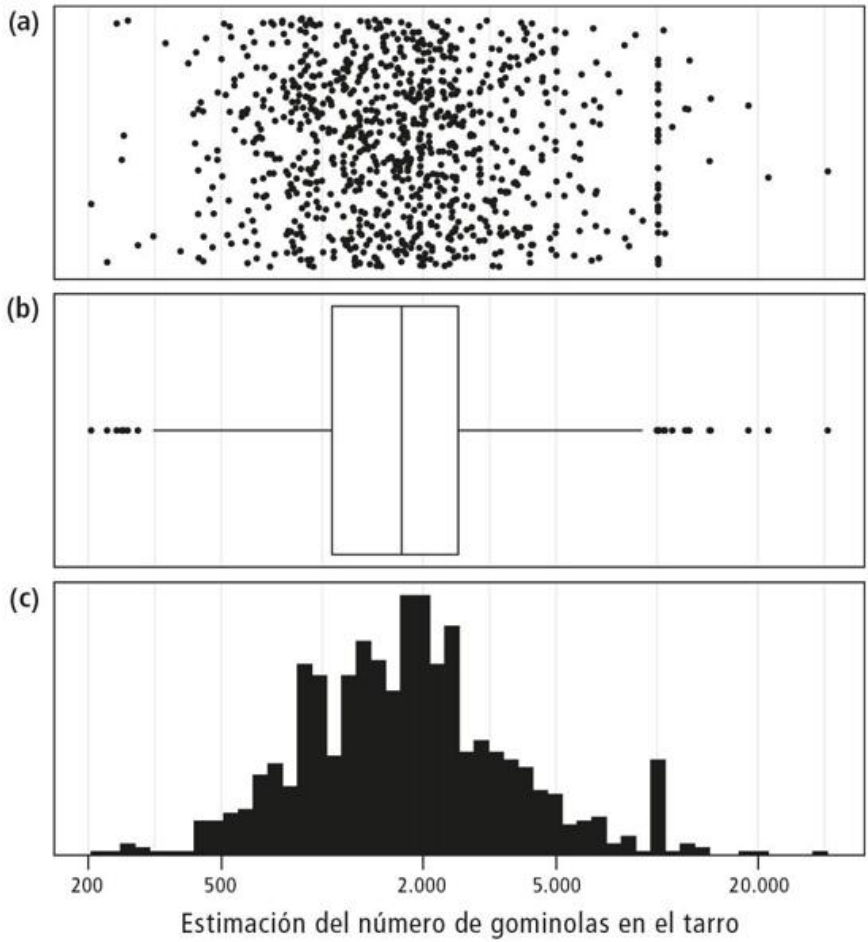


Figura 2.3. Representación gráfica de las estimaciones de gominolas en escala logarítmica. (a) Diagrama de cintas; (b) gráfico de caja y bigotes; (c) histograma; todos muestran una pauta bastante simétrica.

Estas imágenes delatan de manera inmediata algunas características distintivas. La distribución de los datos está muy sesgada, lo que significa que no es ni por asomo simétrica en torno a algún valor central, sino que tiene una larga «cola» a la derecha debida a la ocurrencia de algunos valores muy altos. La serie vertical de puntos en el diagrama de cintas también muestra una cierta preferencia por los números redondos.

Pero hay un problema con todos estos gráficos. La pauta seguida por

los puntos significa que toda la atención se centra en las estimaciones extremadamente altas, con casi todos los números encajonados en el extremo izquierdo del gráfico. ¿Podemos representar los datos de forma más informativa? Podríamos eliminar los valores extremadamente altos por demasiado ridículos (y, de hecho, cuando analizamos originalmente estos datos excluí de manera bastante arbitraria todos los valores superiores a 9.000). También podemos transformar los datos para reducir el impacto de estos valores extremos, por ejemplo, haciendo un gráfico en lo que se denomina escala logarítmica, en la cual el espacio entre 100 y 1.000 es el mismo que el espacio entre 1.000 y 10.000.[30]

La figura 2.3 muestra una pauta más clara, con una distribución bastante simétrica y sin valores atípicos extremos. Esto nos ahorra excluir puntos, que normalmente no es una buena idea, a menos que sean errores claros.

No hay una forma «correcta» de representar conjuntos de números: cada uno de los gráficos que hemos empleado tiene algunas ventajas: los diagramas de dispersión muestran puntos individuales, los gráficos de caja y bigotes son adecuados para lograr un resumen visual rápido y los histogramas son una buena representación visual de la forma subyacente de la distribución de los datos.

Las variables numéricas pueden adoptar distintas formas:

→ Variables contables: En ellas, las mediciones están restringidas a los números enteros 0, 1, 2… Por ejemplo, el número de homicidios diarios, o las estimaciones del número de gominolas en un tarro.

→ Variables continuas: medidas que pueden tener, al menos en principio, una precisión arbitraria. Por ejemplo, la estatura y el peso, cada uno de los cuales puede variar tanto entre individuos como en el tiempo. Por supuesto, pueden ser redondeadas a números enteros de centímetros o kilogramos.

Se denomina generalmente promedio cuando un conjunto de observaciones contables o continuas se reducen a un único dato estadístico de resumen. Todos estamos familiarizados con la idea de, por ejemplo, salarios medios, notas medias o temperaturas medias, pero en ocasiones no está muy claro cómo interpretar estas cifras (particularmente si la persona que cita estas medias no las comprende).

Hay tres interpretaciones básicas del término «promedio», lo que en

ocasiones se conoce como el término «media-mediana-moda»:

→ Media: la suma de los números dividida por el número de casos.

→ Mediana: el valor que está en el medio cuando los números se ponen en orden. Galton, por ejemplo, usó la mediana para resumir los votos en el ejemplo del comienzo de este capítulo.[31]

→ Moda: el valor más común.

Estas tres también se conocen como medidas de tendencia central de la distribución de los datos.

Interpretar el término «promedio» como una media ha generado infinidad de chistes sobre que casi todo el mundo tiene un número de piernas superior a la media (que supuestamente está en torno a 1,99999), y que la gente tiene de media un testículo. Pero no solo con respecto a piernas y testículos puede la media resultar inadecuada. Tanto el número medio de parejas sexuales declaradas como el ingreso medio de un país podrían tener poco que ver con la experiencia de la mayoría de la gente. Esto se debe a que están indebidamente influidos por unos pocos valores extremadamente altos que arrastran al total: [32] pensemos en Warren Beatty o Bill Gates (con respecto, respectivamente, a parejas sexuales e ingresos, obviamente).

La media puede ser muy engañosa cuando los datos en bruto no forman una pauta simétrica alrededor de un valor central, sino que están sesgados hacia un lado, como las estimaciones de gominolas, normalmente con un gran grupo de casos normales pero con una cola de unos pocos valores o muy altos (por ejemplo, los referidos a ingresos) o bajos (por ejemplo, los referidos a piernas). Puedo casi garantizar que, comparado con gente de su edad y sexo, tiene usted un riesgo muy inferior a la media de morir el año que viene. Así, por ejemplo, las tablas de mortalidad del Reino Unido muestran que el 1 % de los hombres de 63 años mueren cada año antes de cumplir los 64, pero muchos de los que morirán ya están gravemente enfermos, y, por ello, la gran mayoría de aquellos que están razonablemente saludables tendrán un riesgo medio menor.

Desgraciadamente, cuando se menciona un «promedio» en los medios de comunicación, en muchas ocasiones no está claro si hay que interpretarlo como una media o como una mediana. Por ejemplo, la Oficina de Estadísticas Nacionales del Reino Unido calcula las ganancias semanales promedio, que es una media, pero también informa de la ganancia semanal mediana por ayuntamiento. En este

caso, sería útil distinguir entre «promedio de ingresos» (media) e «ingreso de la persona promedio» (mediana). Los precios de la vivienda tienen una distribución muy sesgada, con una cola derecha muy larga de propiedades de lujo, y por eso los precios oficiales de la vivienda se expresan como medianas. Pero estas son normalmente presentadas en los medios como el «precio promedio de la vivienda», que es un término muy ambiguo. ¿Se trata del precio de la vivienda promedio (es decir, la mediana)? ¿O del precio promedio de la vivienda (es decir, la media)? Una cosa y otra pueden ser muy diferentes.

Ha llegado el momento de revelar los resultados de nuestro experimento sobre sabiduría de las multitudes con gominolas. No era tan excitante como el peso de un buey, pero pudimos contar con unos pocos votos más que Galton.

Debido a que la distribución de los datos tiene una larga cola a la derecha, la media de 2.408 no es un muy buen resumen de los datos, y la moda de 10.000 parece reflejar más bien una opción extrema en números redondos. Así que probablemente sea mejor seguir a Galton y usar la mediana. Esta resultó ser de 1.775 gominolas. El valor real era de… 1.616.[33] Solo una persona acertó, mientras que el 45 % estimó un número menor y el 55 % un número mayor, lo que indica que no hubo una tendencia sistemática a sobreestimar o subestimar —decimos que el valor real está en el percentil 45 de la distribución de los datos—. La mediana, que está en el percentil 50, sobreestima el valor real en 1.775 − 1.616 = 159, de manera que la mediana sobreestima el valor real en alrededor de un 10 %, y solo aproximadamente 1 de cada 10 personas se aproximó tanto. De manera que la sabiduría de las multitudes fue bastante buena, se acercó más a la verdad que el 90 % de la gente.

Estadísticos descriptivos de las estimaciones	Datos completos
Media	2.408
Mediana	1.775
Moda	10.000
Rango	219 a 31.337
Rango intercuartílico	1.109 a 2.599
Desviación típica	2.422

Tabla 2.1. Estadísticos descriptivos para las 915 estimaciones sobre el número de gominolas. El número correcto era 1.616.

Cómo describir la dispersión de una distribución de datos

No basta con dar un único valor resumen de una distribución: necesitamos tener una idea de su dispersión, lo que en ocasiones se conoce como su variabilidad. Por ejemplo, conocer la talla promedio de los zapatos de un hombre adulto no servirá de mucho a una empresa de calzado cuando tenga que decidir cuántos zapatos fabricar de cada talla. Una talla no le sirve a todo el mundo, algo que se hace patente, por ejemplo, en los asientos para pasajeros de los aviones.

La tabla 2.1 nos muestra una variedad de estadísticos descriptivos para las estimaciones del número de gominolas, incluyendo tres formas de resumir la dispersión. El rango es una buena opción, pero claramente es muy sensible a valores extremos, como la estimación aparentemente extraña de 31.337 gominolas.[34] Por el contrario, el rango intercuartílico (RIC) no se ve afectado por los valores extremos. Es la distancia entre los percentiles 25 y 75 de los datos, y contiene, por tanto, la «mitad central» de los números, en este caso, entre 1.109 y 2.599 gominolas; la «caja» central de los gráficos de caja y bigotes mostrados más arriba cubriría el rango intercuartílico. Finalmente, la desviación típica es una medida de dispersión ampliamente utilizada. Es la medida técnicamente más compleja, pero es solo realmente apropiada para datos simétricos,[35] dado que se ve indebidamente afectada por valores atípicos. Por ejemplo, si eliminamos el valor 31.337 de los datos, la desviación típica se reduce de 2.422 a 1.398. [36]

Las «multitudes» de nuestro pequeño experimento demostraron tener una sabiduría considerable, a pesar de algunas respuestas algo raras. Esto demuestra que los datos a menudo contienen algunos errores y valores atípicos, pero no necesariamente hay que identificar y excluir cada uno de ellos. También muestra los beneficios de usar medidas descriptivas que no se vean indebidamente afectadas por observaciones raras como ese 31.337 —las que se conocen como medidas robustas, que incluyen la mediana y el rango intercuartílico —. Finalmente, muestra lo útil que resulta echar un simple vistazo a los datos, una lección que remacharemos con el siguiente ejemplo.

Cómo describir las diferencias entre grupos de números

¿Cuántas parejas sexuales afirman haber tenido a lo largo de su vida las personas que viven en Gran Bretaña?

Esta pregunta no pretende solo cotillear en la vida privada de la gente. Cuando el sida alcanzó proporciones preocupantes en la década de

1980, los responsables de salud pública se dieron cuenta de que no había ninguna evidencia sólida sobre el comportamiento sexual en Gran Bretaña, particularmente sobre la frecuencia con la cual las personas cambiaban de pareja, cuántas personas tenían múltiples parejas simultáneamente y cuáles eran sus prácticas sexuales. Este conocimiento era esencial para predecir la expansión de enfermedades de transmisión sexual y para planificar la respuesta sanitaria, y, a pesar de ello, aún se seguían empleando los poco fiables datos recopilados por Alfred Kinsey en Estados Unidos en la década de 1940, que en ningún momento pretendían constituir una muestra representativa.

Así que, a partir de finales de la década de 1980, se llevaron a cabo encuestas grandes y costosas sobre comportamiento sexual en el Reino Unido y Estados Unidos, a pesar de que desde algunos sectores se produjo una fuerte oposición a estas. En el Reino Unido, Margaret Thatcher retiró en el último minuto la financiación a una gran encuesta sobre hábitos sexuales. No obstante, sus organizadores tuvieron la suerte de encontrar financiación privada, y el resultado fue la Encuesta Nacional de Actitudes y Hábitos Sexuales (Natsal),[37] que se lleva a cabo en el Reino Unido cada diez años desde 1990.

La tercera encuesta, conocida como Natsal-3, fue llevada a cabo en 2010 y costó siete millones de libras.[38] La tabla 2.2 muestra los estadísticos descriptivos referidos al número de parejas sexuales (de sexo opuesto) declarado por individuos de 35 a 44 años según la encuesta Natsal-3. Siempre es un ejercicio interesante usar exclusivamente estos descriptivos para intentar reconstruir cuál es la pauta de los datos. Se aprecia que el valor más común (la moda) es 1, referido a las personas que solo han tenido una pareja a lo largo de su vida, y, aun así, el rango es enorme. Esto se refleja también en la sustancial diferencia existente entre las medias y las medianas, lo que siempre es una señal de que la distribución de los datos tiene una larga cola derecha. La desviación típica es alta, pero se trata de una medida inapropiada de dispersión para una distribución de los datos de este tipo, dado que se ve indebidamente influida por unos pocos valores extremadamente altos.

Número declarado de parejas sexuales a lo largo de la vida	Hombres, edad 35-44	Mujeres, edad 35-44
Media	14,3	8,5
Mediana	8	5
Moda	1	1
Rango	0 a 500	0 a 550
Rango intercuartílico	4 a 18	3 a 10
Desviación típica	24,2	19,7

Tabla 2.2. Estadísticos descriptivos del número de parejas sexuales (de sexo opuesto) a lo largo de la vida, declaradas por 806 hombres y 1.215 mujeres de 35 a 44 años, basados en entrevistas llevadas a cabo en Natsal-3 entre 2010 y 2012. Se incluyen las desviaciones típicas, aunque son resúmenes inapropiados de la distribución de esos datos.

Si comparamos las respuestas de los hombres y de las mujeres, veremos que los hombres afirman haber tenido de media 6 parejas sexuales más que las mujeres, o, de otra manera, que el hombre promedio (el mediano) declara haber tenido 3 parejas sexuales más que la mujer promedio. O que, en términos relativos, los hombres declaran haber tenido un 60 % más de parejas que las mujeres, tanto para el caso de la media como de la mediana.

Esta diferencia podría generar sospechas sobre los datos. ¡En una población cerrada con el mismo número de hombres y mujeres con una edad similar, es un hecho matemático que el número medio de parejas sexuales del sexo opuesto debería ser esencialmente el mismo para hombres y mujeres![39] Entonces, ¿por qué los hombres de este grupo de edad, de 35 a 44, declaran muchas más parejas que las mujeres? En parte se puede deber a hombres que tengan parejas más jóvenes, pero también a que parece haber diferencias sistemáticas en la manera en la que hombres y mujeres cuentan sus historias sexuales e informan de estas. Podemos sospechar que es más probable que los hombres exageren el número de sus parejas, o que las mujeres lo minimicen, o ambas cosas.

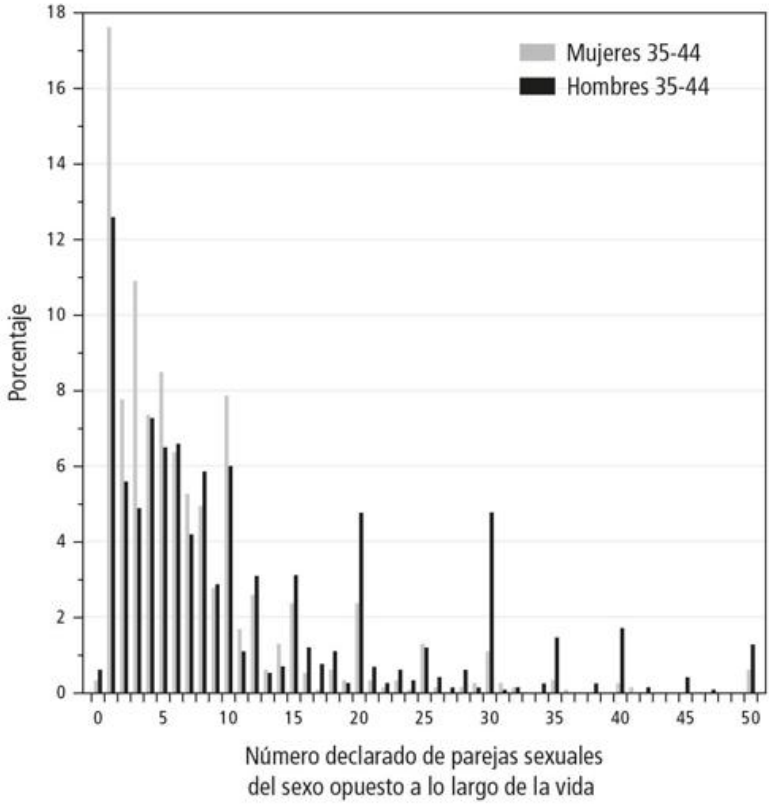


Figura 2.4. Datos derivados de Natsal-3 basados en entrevistas entre 2010 y 2012. La serie ha sido truncada en 50 por razones de espacio —los números totales van hasta 500 tanto para hombres como para mujeres—. Nótese cómo los entrevistados han usado números redondos para 10 o más parejas, y la tendencia de los hombres a declarar más parejas que las mujeres.

La figura 2.4 revela la distribución real de los datos, que apoya la impresión dada por la estadística descriptiva de una larga cola derecha de valores extremos. Pero solo echando un vistazo a los datos en bruto descubrimos otros detalles importantes, como la fuerte tendencia tanto entre los hombres como entre las mujeres a proporcionar números redondos cuando han tenido diez o más parejas (excepto el caso de un hombre bastante puntilloso, posiblemente un estadístico, que dijo que había tenido exactamente «cuarenta y siete»). Se podría, por supuesto, dudar de la fiabilidad de todas estas declaraciones, y de hecho, en el siguiente capítulo discutiremos los sesgos potenciales de estos datos.

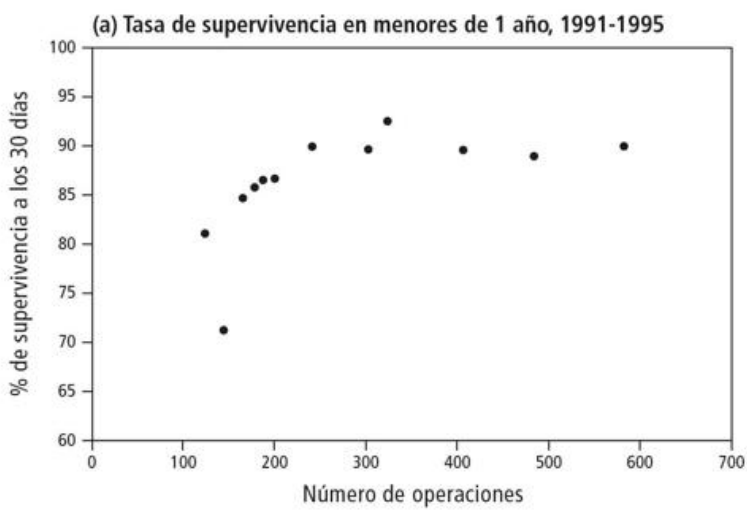
Cotidianamente se reúnen y se comunican grandes conjuntos de datos

numéricos utilizando unos pocos estadísticos de tendencia central y de dispersión; tal como muestra el ejemplo de las parejas sexuales, estos estadísticos nos proporcionan mucha información acerca de la pauta global de los datos. Por tanto, hay que saber observar los datos cuidadosamente. El siguiente ejemplo muestra que una buena visualización es particularmente valiosa cuando queremos identificar la pauta que hay detrás de un conjunto grande y complejo de números.

Cómo describir las relaciones entre variables

¿Tienen los hospitales más ajetreados unas tasas de supervivencia más altas?

Hay un considerable interés en el llamado «efecto de volumen» en la cirugía: la idea de que los hospitales más ajetreados tienen mejores tasas de supervivencia, posiblemente porque alcanzan una mayor eficiencia y una mayor experiencia. La figura 2.5 muestra las tasas de supervivencia a 30 días en hospitales del Reino Unido que llevan a cabo cirugía cardiaca infantil cruzadas con el número de niños tratados. La figura 2.5(a) muestra los datos para niños menores de 1 año a lo largo del período 1991-1995 que presentamos al comienzo del anterior capítulo, puesto que este grupo de edad tiene un riesgo más alto y fue el objeto de la investigación de Bristol. La figura 2.5(b) muestra los datos para niños de menos de 16 años en el período 2012-2015 que presentamos previamente en la tabla 1.1 —los datos específicos para los niños menores de 1 año no están disponibles para ese período—. El volumen está representado en el eje horizontal x, y la tasa de supervivencia en el eje vertical y.[40]



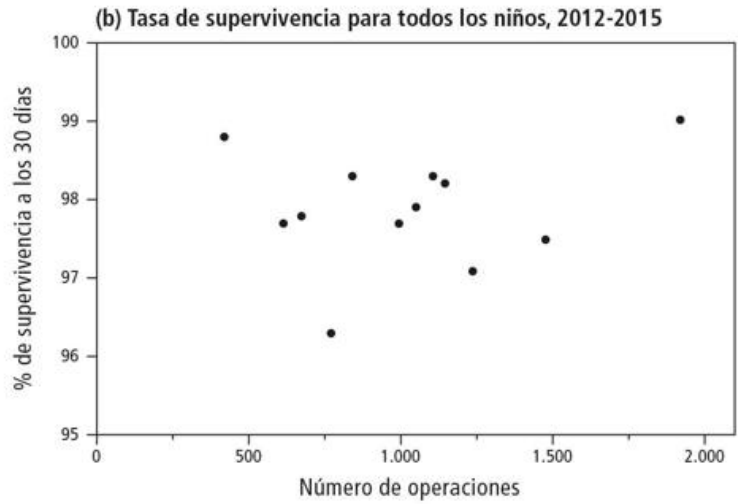


Figura 2.5. Diagramas de dispersión de las tasas de supervivencia cruzadas con el número de operaciones de cirugía cardiaca infantil. Para (a) 1991-1995 la correlación de Pearson es de 0,59 y la correlación de rango es de 0,85, para (b) 2012-2015, la correlación de Pearson es de 0,17 y la correlación de rango es de –0,03.

Los datos de 1991-1995 de la figura 2.5(a) incluyen un valor atípico claro, un pequeño hospital con solo un 71 % de tasa de supervivencia. Se trata del de Bristol, cuyas bajas tasas de supervivencia y la investigación pública subsiguiente que estas provocaron fueron tratadas en el capítulo 1. Pero incluso eliminando Bristol (intente poner el pulgar sobre ese valor atípico) la pauta de los datos para 1991-1995 sugiere que hay tasas de supervivencia mayores en aquellos hospitales que llevan a cabo más operaciones.

Resulta útil utilizar un número único para resumir una relación creciente o decreciente entre las parejas de números que se muestran en un diagrama de dispersión. Normalmente para ello se usa el coeficiente de correlación de Pearson, una idea propuesta originalmente por Francis Galton pero publicada formalmente en 1895 por Karl Pearson, uno de los fundadores de la moderna estadística. [41]

Una correlación de Pearson va de -1 a 1, y representa la cercanía a una línea recta de los puntos de datos. Una correlación de 1 se da cuando todos los puntos se encuentran en una línea recta creciente, mientras que una correlación de -1 se da cuando todos los puntos se encuentran en una línea recta decreciente. Una correlación cercana a 0 puede ser el resultado de una dispersión aleatoria de puntos, o

cualquier otra pauta en la cual no hay ninguna tendencia sistemática creciente o decreciente, algunos ejemplos de las cuales se muestran en la figura 2.6.

La correlación de Pearson es de 0,59 para los datos de 1991-1995 que se muestran en la figura 2.5(a), lo que sugiere una asociación entre el volumen y la supervivencia. Si Bristol se elimina, la correlación de Pearson se incrementa hasta 0,67, dado que los puntos restantes están más alineados en una línea recta. Una medida alternativa es la correlación de rango de Spearman, llamada así por el psicólogo inglés Charles Spearman (que desarrolló la idea de la inteligencia general subyacente), y que depende solo de los rangos de los datos, y no de sus valores específicos. Esto significa que puede ser cercana a 1 o -1 si los puntos están más cerca de una línea creciente o decreciente, incluso si esta línea no es recta; la correlación de rango de Spearman para los datos de la figura 2.5(a) es 0,85, considerablemente mayor que la correlación de Pearson, dado que los puntos están más cerca de una curva creciente que de una línea recta.

Figura 2.6. Dos conjuntos de puntos de datos (ficticios) para los cuales

los coeficientes de correlación de Pearson son 0. Esto claramente no significa que no haya ninguna relación entre las dos variables representadas en los gráficos. La figura inferior es el maravilloso Datasaurus Dozen. (Los datos son del «Datasaurus Dozen», https://www.autodeskresearch.com/publications/samestats, incluyendo A. Cairo, «Download the Datasaurus: Never Trust Summary Statistics Alone: Always Visualize Your Data», http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html).

La correlación de Pearson es 0,17 para los datos de 2012-2015 en la figura 2.5(b), y la correlación de rango de Spearman es −0,03, lo que sugiere que ha dejado de haber una relación clara entre el número de casos y las tasas de supervivencia. No obstante, con un número tan bajo de hospitales el coeficiente de correlación puede ser muy sensible a los datos individuales —si eliminamos el hospital más pequeño, que tiene una alta tasa de supervivencia, la correlación de Pearson salta hasta 0,42—.

Los coeficientes de correlación son simples resúmenes de asociaciones, y no pueden usarse para concluir con seguridad que hay una relación subyacente entre volumen y tasas de supervivencia, por no hablar de por qué podría haberla.[42] En muchas aplicaciones, el eje x representa lo que se conoce como la variable independiente, y lo que nos interesa es su influencia en la variable dependiente del eje y. Pero, como veremos con más detalle en el capítulo 4 sobre causalidad, esto presupone la dirección en la cual podría darse esta influencia de una variable sobre otra. Incluso en la figura 2.5(a) no podemos concluir que las tasas más altas de supervivencia hayan sido causadas por el aumento del número de casos —de hecho, podría incluso ser al revés: los mejores hospitales simplemente son los que atraen más pacientes —.

Cómo describir tendencias

¿Cuál es la pauta de crecimiento de la población mundial en el último medio siglo?

La población mundial está aumentando, y la comprensión de los factores que explican el cambio demográfico es de una importancia capital para prepararnos para los desafíos a los que los distintos países se enfrentan en la actualidad y se enfrentarán en el futuro. La División de Población de las Naciones Unidas genera estimaciones de población para todos los países del mundo desde 1951, junto con proyecciones hasta 2100.[43] A continuación nos ocuparemos de las tendencias mundiales desde 1951.

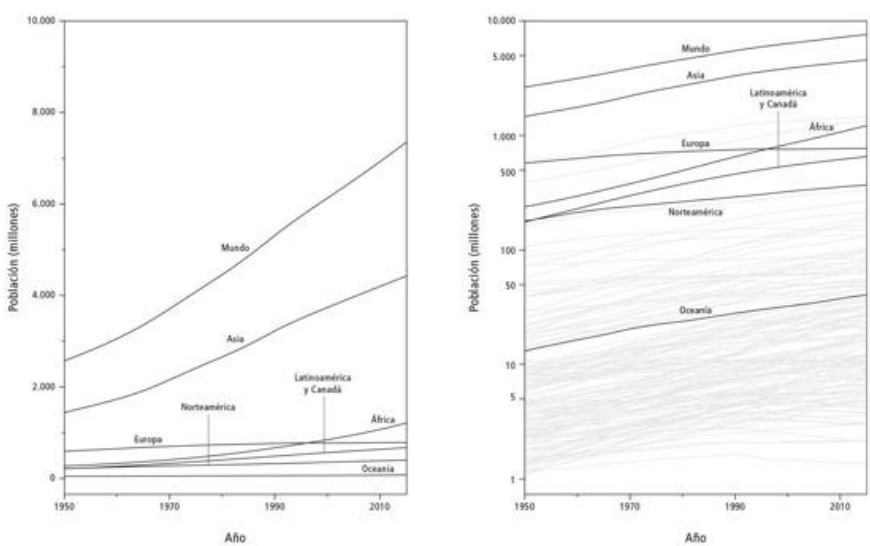


Figura 2.7. Población total mundial, de los continentes y de los países entre 1950 y 2015, incluyendo ambos sexos: (a) muestra tendencias en una escala normal, (b) en escala logarítmica, junto con las líneas de tendencia para países individuales que tenían una población de al menos un millón de habitantes en 1951.

La figura 2.7(a) es un gráfico simple de líneas que refleja la evolución de la población mundial desde 1951, que, como vemos, se ha triplicado durante ese período, hasta alcanzar casi los 7.500 millones. El aumento se ha debido en gran medida a los países asiáticos, mientras que en el caso del resto de los continentes es difícil distinguir una pauta. Por el contrario, la escala logarítmica de la figura 2.7(b) permite separar las tendencias seguidas por los continentes, revelando un crecimiento más pronunciado en África, y una tendencia más plana en otros continentes, especialmente Europa, donde la población ha empezado a disminuir recientemente.

Las líneas grises de la figura 2.7(b) representan los cambios en países individuales, pero es imposible distinguir desviaciones de la tendencia general creciente. La figura 2.8 utiliza un resumen simple de la tendencia de cada país —el incremento relativo entre 1951 y 2015—. Un incremento relativo de 4 significa que hay cuatro veces más población en 2015 que en 1951 (es el caso, por ejemplo, de Liberia, Madagascar y Camerún). Que los símbolos sean proporcionales al tamaño del país hace que todos los ojos se posen en los países más grandes, y agrupar los países por continentes nos permite detectar inmediatamente tanto grupos de países como casos atípicos. Siempre es útil separar los datos según un factor —en este caso los continentes — que explique parte de la variabilidad global.

El gran crecimiento demográfico de África se mantiene, pero con una amplia variación, y un caso extremo, Costa de Marfil. Asia también muestra una enorme variación, lo que refleja la amplia diversidad de países en ese continente, con Japón y Georgia en un extremo y Arabia Saudita, que experimentó el mayor crecimiento del mundo entero, en el otro. El incremento en Europa ha sido relativamente bajo.

Como cualquier buen gráfico, este plantea más preguntas y anima a seguir investigando, tanto en identificar países individuales como, por supuesto, en examinar cómo serán las tendencias en el futuro.

Hay claramente muchísimas maneras de examinar una base de datos tan compleja como las cifras demográficas de la ONU, y ninguna de ellas puede considerarse la «correcta». No obstante, Alberto Cairo ha identificado cuatro características comunes de una buena

visualización de los datos:

1. Contiene información fiable.

2. El diseño se ha escogido para destacar las pautas más relevantes.

3. La presentación es atractiva, pero las apariencias no pueden ser un sustituto de la honestidad, la claridad y la profundidad.

4. Cuando resulte apropiado, está organizada de tal manera que permita una cierta exploración de los datos.

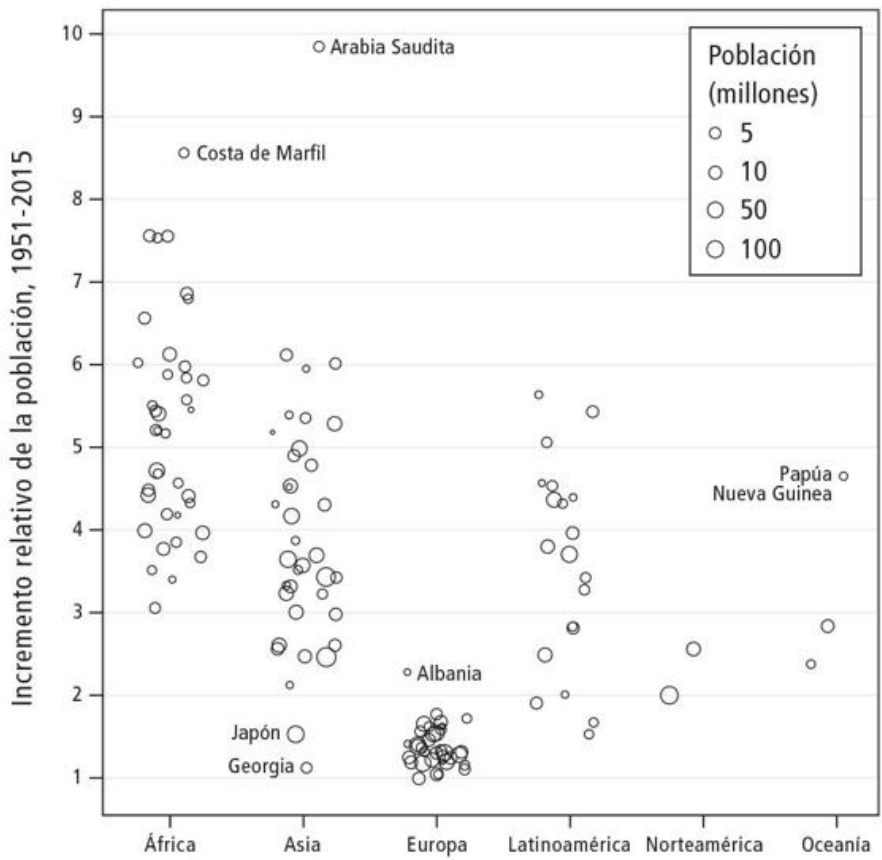


Figura 2.8. Incremento de la población entre 1951 y 2015 para países con al menos un millón de habitantes en 1951.

La última de estas características se puede conseguir permitiendo que se interactúe con la visualización, y aunque esto es difícil de ilustrar en un libro, el siguiente ejemplo muestra las ventajas de personalizar un gráfico.

¿Cuán popular ha sido mi nombre a lo largo del tiempo?

Algunos gráficos son tan complejos que es difícil apreciar pautas interesantes a primera vista. Tomemos por ejemplo la figura 2.9, en la cual cada línea muestra la evolución de la popularidad de un determinado nombre de chico para los nacidos en Inglaterra y Gales entre 1905 y 2016.[44] Se trata de una historia social extraordinaria, y, no obstante, por sí sola únicamente nos informa de las modas rápidamente cambiantes en cuanto a nombres se refiere, y de una mayor amplitud y diversidad de nombres desde mediados de la década de 1990, como sugieren las líneas más recientes y densas.

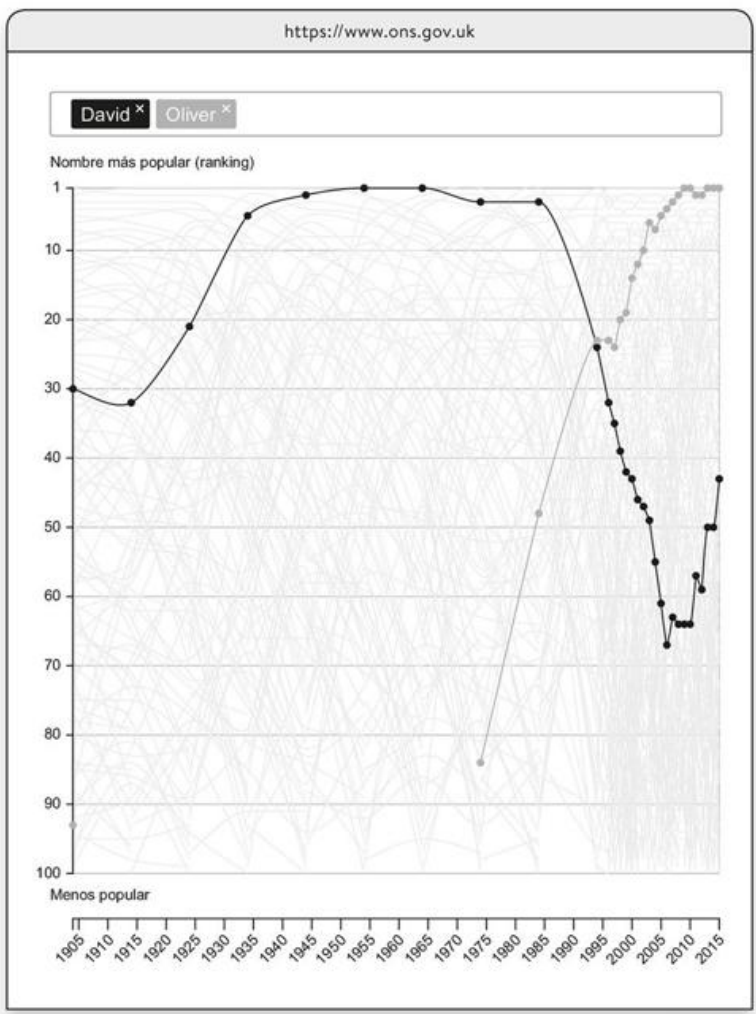


Figura 2.9. Captura de pantalla de un gráfico interactivo proporcionado por la Oficina de Estadísticas Nacionales del Reino

Unido que muestra la tendencia de cada nombre de chico en una liga de popularidad. Mis singularmente poco imaginativos padres me dieron el nombre de chico más popular en 1953, pero desde entonces he pasado de moda, todo lo contrario que le ha pasado a Oliver. David, no obstante, ha mostrado señales recientes de recuperación, posiblemente influido por David Beckham.

Solo permitiendo la interactividad podemos apreciar líneas específicas que tengan un interés personal. Yo, por ejemplo, tengo curiosidad por la tendencia seguida por el nombre «David», un nombre que se hizo especialmente popular en las décadas de 1920 y 1930 debido a que el príncipe de Gales (posteriormente el rey breve Eduardo VIII) se llamaba David. Pero su popularidad ha decaído precipitadamente: en 1953 yo era uno de las decenas de miles de niños con el nombre de David, pero en 2016 solo 1.461 recibieron ese nombre, y alrededor de otros 40 nombres eran más populares.

Comunicación

Este capítulo se ha centrado en el resumen y la comunicación de los datos de forma abierta y no manipuladora; no queremos influir en las emociones y actitudes de nuestra audiencia, o convencerlos de una determinada perspectiva. Solo queremos contar cómo son las cosas, o al menos cómo parecen ser, y aunque no podemos decir que lo que contamos sea la verdad absoluta, al menos podemos intentar ser lo más sinceros posible.

Por supuesto, lograr la objetividad científica es más difícil de lo que parece. Cuando la Sociedad Estadística de Londres (lo que más tarde sería la Real Sociedad Estadística) fue creada, en 1834, por Charles Babbage, Thomas Malthus y otros, declararon solemnemente que «la Sociedad Estadística considera que la primera y más importante regla de conducta es la exclusión de toda opinión de sus transacciones y publicaciones —confinar su atención vigorosamente a los hechos— y, en la medida de lo posible, a hechos que puedan ser representados numéricamente y presentados en tablas».[45] Luego hicieron caso omiso de este principio, e inmediatamente empezaron a introducir sus opiniones sobre el significado de sus datos sobre criminalidad, salud y economía, y sobre qué cabía hacer al respecto. Quizá lo mejor que podamos hacer es reconocer que esta tentación existe, y hacer lo que podamos para guardarnos nuestras opiniones para nosotros mismos.

La primera regla de la comunicación es callarse y escuchar, para poder así conocer algo sobre la audiencia a la que estás comunicando algo, ya sea esta políticos, profesionales o el público en general. Tenemos

que comprender sus inevitables limitaciones y la posibilidad de malentendidos, y combatir la tentación de ser demasiado sofisticados y sabelotodos, o de presentar las cosas con demasiado detalle.

La segunda regla de la comunicación es saber lo que quieres conseguir. Idealmente, el objetivo es favorecer un debate abierto, y una toma de decisiones informada. Pero no está de más recordar de nuevo que los números no hablan por sí mismos; el contexto, el lenguaje y el diseño gráfico determinan la forma en la que se recibe la información. Tenemos que reconocer que estamos contando una historia, y es inevitable que la gente haga comparaciones y juicios, por mucho que queramos informar, no persuadir. Todo lo que podemos hacer es intentar evitar reacciones viscerales, diseñando nuestra comunicación con ese fin.

Cómo contar historias a través de la estadística

Este capítulo ha introducido el concepto de visualización de los datos, lo que normalmente se conoce como «dataviz». Estas técnicas tienen como destinatarios fundamentalmente investigadores o audiencias bastante sofisticadas, y emplean toda una colección de gráficos seleccionados por su valor en la comprensión y exploración de los datos, más que por lo vistosos que sean. Cuando hemos obtenido de los datos el mensaje que queremos comunicar, podemos usar la infografía, o «infoviz», para atraer la atención de la audiencia y contar una historia interesante.

Podemos ver regularmente en los medios ejemplos de infografías sofisticadas, pero la figura 2.10 nos muestra un ejemplo muy sencillo, en el que se cuenta de manera bastante rigurosa la evolución de una serie de tendencias sociales basándose en las respuestas a tres preguntas de la Encuesta Nacional del Reino Unido de Actitudes y Estilos de Vida Sexuales (Natsal-3) de 2010: ¿a qué edad los hombres y las mujeres tuvieron por primera vez relaciones sexuales?, ¿a qué edad comenzaron a cohabitar por primera vez?, y ¿a qué edad tuvieron su primer hijo?[46] Las edades medianas para cada uno de estos eventos están cruzadas por el año de nacimiento de las mujeres, y los tres puntos están conectados por una línea gruesa vertical. El incremento de la longitud de esta línea entre las mujeres nacidas en la década de 1930 y las nacidas en la de 1970 es un reflejo de que durante un período cada vez mayor de la vida de la mujer es necesario usar métodos anticonceptivos con anterioridad a tener el primer hijo.

Figura 2.10. Infografía basada en datos de la Encuesta Nacional del Reino Unido sobre Actitudes y Estilos de Vida Sexuales (Natsal-3); las lecciones que extraer de los datos se indican visual y verbalmente.

Los gráficos dinámicos son aún más avanzados; en ellos se usa el movimiento para revelar pautas en los cambios producidos a lo largo del tiempo. El auténtico maestro de esta técnica fue Hans Rosling, cuyos vídeos y conferencias TED eran el modelo en todo lo relativo a comunicar datos estadísticos. Un ejemplo de ello era la representación de la relación entre los cambios en la riqueza y en la salud a través del movimiento animado de burbujas que representaban el progreso de cada país desde 1800 hasta la actualidad. Rosling usó sus gráficos para intentar corregir malentendidos acerca de la distinción entre países «desarrollados» y «subdesarrollados». Los gráficos dinámicos

revelaron, a lo largo del tiempo, que casi todos los países se movieron de acuerdo con una pauta común hacia una mayor salud y prosperidad.[47]

En este capítulo hemos ido desde las simples descripciones y los gráficos de datos en bruto hasta los ejemplos complejos de cómo contar historias con la estadística. La moderna informática ha hecho que la visualización de los datos sea cada vez más sencilla y más flexible; y dado que los estadísticos descriptivos pueden esconder tanto como iluminar, es esencial una representación gráfica apropiada. No obstante, resumir y comunicar números en bruto es solo el primer paso en el proceso de aprender de los datos. Para avanzar en este camino, necesitamos aclarar, para empezar, qué es lo que estamos intentando conseguir.

Resumen

→ Se pueden usar toda una serie de estadísticos para resumir la distribución empírica de los datos, incluyendo medidas de posición y de dispersión.

→ Los datos sesgados son algo normal, y algunos estadísticos de resumen son muy sensibles a los valores atípicos.

→ Los resúmenes de datos siempre esconden algún detalle, por lo que se necesita tener cuidado para que no se pierda información.

→ Los conjuntos de números pueden ser visualizados en diagramas de cintas, gráficos de caja y bigotes e histogramas.

→ Puede valer la pena realizar transformaciones para revelar mejor las pautas, y es aconsejable echar un vistazo para detectar pautas, valores atípicos, similitudes y agrupaciones de datos.

→ Las parejas de números conviene representarlas como diagramas de dispersión, y las series temporales, como gráficos de línea.

→ Cuando se exploran datos, un objetivo primordial es encontrar factores que expliquen la variación global.

→ Los gráficos pueden ser interactivos y animados.

→ La infografía sirve para destacar características interesantes y puede guiar al espectador a través de una historia, pero se debe ser consciente de su objetivo y de su impacto.

[26] La eugenesia es la idea de que la raza humana puede ser mejorada mediante la cría selectiva, o bien favoreciendo que los más «aptos» produzcan más hijos, por ejemplo mediante incentivos financieros, o evitando que los «no aptos» se reproduzcan, por ejemplo, mediante su esterilización. Muchos de los pioneros en el desarrollo de la estadística eran defensores entusiastas de la eugenesia. La experiencia de la Alemania nazi puso fin al movimiento, aunque la revista académica Annals of Eugenics no cambió su nombre por el de Annals of Genetics hasta 1955.

[27] F. Galton, «Vox Populi», Nature, 1907, disponible en https://www.nature.com/articles/075450a0.

[28] El término «distribución» es empleado con mucha frecuencia en estadística y puede ser ambiguo, por lo que intentaré aclarar qué significa en cada caso. Los gráficos se han llevado a cabo por medio del software libre R.

[29] En esta versión concreta del diagrama de caja y bigotes, la barra central en negrita representa la mediana (el punto intermedio), la caja contiene la mitad central de los puntos, mientras que los «bigotes» muestran los valores más bajos y más altos, aparte de los valores atípicos, que se representan individualmente.

[30] Para obtener el logaritmo de un número x, encontramos la potencia de 10 que nos dé x, de manera que, por ejemplo, el logaritmo de 1.000 es 3, dado que 103 = 1.000. Las transformaciones logarítmicas son particularmente apropiadas cuando es razonable asumir que la gente está cometiendo errores «relativos» más que «absolutos», por ejemplo, porque esperaríamos que la gente se equivocase, digamos, por un factor relativo, digamos, del 20 % en cada dirección, en lugar de, digamos, 200 gominolas de diferencia con respecto al número correcto, con independencia de si la estimación es un valor alto o bajo.

[31] Aunque en una carta enviada a Nature alguien cuestionó que Galton hubiese empleado la mediana, afirmando que la media habría proporcionado una estimación más cercana al número real.

[32] Imagínese tres personas en una habitación cuyos ingresos mensuales son 400, 500 y 600 libras, de manera que su ingreso medio sería 1.500 / 3 = 500 libras, una cantidad igual a la mediana. Entonces entran 2 personas que ganan 5.000 libras a la semana: el ingreso medio se dispara a 11.500 / 5 = 2.300 libras, mientras que la mediana apenas se mueve, quedándose en 600 libras.

[33] En el vídeo (https://www.youtube.com/watch?v=n98BhnwWmsc) de nuestro experimento eliminé de manera bastante arbitraria las 33 estimaciones más altas, de 9.999 o más, saqué el logaritmo para tener una agradable distribución simétrica, tomé la media de esta distribución transformada y la volví a transformar para obtener la estimación en la escala original. Esto me dio la cifra de 1.680 como la «mejor estimación», que resultó ser la más cercana de todas las estimaciones al valor real de 1.616. Este proceso —tomar el logaritmo, calcular la media, volver a la escala original— lleva a lo que se conoce como la media geométrica. Esta es equivalente a multiplicar todos los números juntos, y, si hay n números, calcular su raíz enésima.

La media geométrica se emplea en la creación de algunos índices económicos, particularmente los basados en razones. Esto es debido a que tiene la ventaja de que no importa cuál es la dirección de la razón específica: el coste de las naranjas puede medirse en libras por naranja o en naranjas por libras, y al final siempre tendrían la misma media geométrica, mientras que escoger una opción u otra podría suponer una gran diferencia si empleamos la media como promedio.

[34] Seguramente fue una errata, la estimación real debía ser más bien 1.337, que es la forma numérica de denominar el término «leet», que en jerga de internet se emplea para referirse a alguien especialmente apto. Había 9 estimaciones de 1.337.

[35] El índice de Gini es una medida de dispersión para datos muy sesgados, como los ingresos, y es ampliamente usado como medida de desigualdad, pero tiene una forma compleja y contraintuitiva.

[36] El cuadrado de la desviación típica se conoce como varianza. Es difícil de interpretar directamente, pero es útil desde un punto de vista matemático. El valor «31.337» fue probablemente una opción deliberada por parte de un friki mayorcito, relacionada con una forma antigua de referirse a «eleet / elite».

[37] Por su nombre en inglés, National Sexual Attitudes and Lifestyle Survey. (N. del T.).

[38] C. H. Mercer *et al.*, «Changes in Sexual Attitudes and Lifestyles in Britain through the Life Course and Over Time: Findings from the National Surveys of Sexual Attitudes and Lifestyles (Natsal)», The Lancet 382 (2013), pp. 1781-1794. Para un examen brillante de las estadísticas sexuales, véase D. Spiegelhalter, Sex by Numbers, Wellcome Collection, 2015.

[39] Esto se debe a que el conjunto de todos los hombres, y el de todas las mujeres, tienen el mismo número total de parejas, dado que cada pareja está formada por un hombre y una mujer. De manera que si los grupos son del mismo tamaño, las medias deben ser las mismas. Cuando discuto este tema en los colegios, uso el ejemplo de las parejas de baile o de los apretones de manos.

[40] Aunque las tasas de supervivencia globales en las dos figuras no son directamente comparables, dado que los niños son de distintos rangos de edad, de hecho la supervivencia para niños de todas las edades se ha incrementado del 92 % al 98 % a lo largo de estos veinte años.

[41] Karl Pearson era un gran entusiasta de todo lo alemán: incluso se cambió la grafía de su nombre de Carl a Karl, aunque esto no le impidió aplicar sus conocimientos estadísticos a la balística en la Primera Guerra Mundial. En 1911 fundó el primer Departamento de Estadística del mundo en el University College de Londres, y ocupó la Galton Chair de Eugenesia, financiada a través del legado de Francis Galton.

[42] Las tasas de supervivencia se basan en distintos números de casos y, por tanto, están sujetas a distintos grados de variabilidad aleatoria. De manera que, aunque una correlación siempre se puede considerar como una descripción de un conjunto de datos, cualquier inferencia formal tiene que tener en cuenta el hecho de que los datos son proporciones. Mostraré cómo hacerlo en el capítulo 6.

[43] https://esa.un.org/unpd/wpp/Download/Standard/Population/.

[44] Los datos de la Oficina de Estadísticas Nacionales sobre popularidad de los nombres están disponibles en https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/babynamesenglandandwales/2015.

[45] I. D. Hill, «Statistical Society of London-Royal Statistical Society: The First 100 Years: 1834-1934», Journal of the Royal Statistical Society: Series A (General) 147: 2 (1984), pp. 130-139.

[46] https://www.natsal.ac.uk/sites/default/files/2021-04/Natsal-3%20infographics%20%281%29_0.pdf.

[47] Por desgracia, un libro estático con gráficos en blanco y negro no es el medio más adecuado para presentar su trabajo, por lo que recomiendo que eche un vistazo a gapminder.org. Rosling participó en

un debate en televisión con un periodista danés que estaba soltando el tipo de ideas erróneas que Hans había pasado su vida entera intentando contrarrestar. Hans simplemente replicó que «estos hechos no admiten discusión. Yo tengo razón, y usted está equivocado» —que, para un estadístico, es un ejemplo raro de hablar claro—. Para más información sobre el trabajo de Rosling, véase H. Rosling, Unveiling the Beauty of Statistics for a Fact-Based World View, disponible en www.gapminder.org.

03

Pero, en realidad,

¿por qué nos interesan los datos?

Poblaciones y mediciones

¿Cuántas parejas sexuales han tenido realmente las personas que viven en Gran Bretaña?

El capítulo anterior mostró algunos resultados de lo más interesantes provenientes de una encuesta británica reciente en la cual la gente declaraba cuál había sido el número de parejas sexuales que había tenido a lo largo de su vida. El gráfico resultante reveló varias características, incluida una cola (muy) larga, una tendencia a usar números redondos como 10 o 20 y que los hombres declaran más parejas que las mujeres. Pero los investigadores que habían gastado millones de libras para reunir estos datos no estaban realmente interesados en lo que decían estos entrevistados en concreto —después de todo, se les había garantizado su total anonimato—. Sus respuestas eran un medio para un fin: decir algo sobre la pauta global de las relaciones sexuales en Gran Bretaña (las de los millones de personas que no fueron entrevistados sobre su comportamiento sexual).

Extraer conclusiones sobre toda Gran Bretaña a partir de las respuestas a una encuesta no es algo trivial. En realidad, esto no es así: es increíblemente fácil afirmar que lo que han dicho esos entrevistados representa fielmente lo que pasa realmente en el país. Y eso es lo que hacen continuamente las encuestas de los medios sobre sexo, en las que la gente rellena voluntariamente formularios en páginas web sobre lo que dicen que hacen en la intimidad.

El proceso de ir de las respuestas en bruto de la encuesta a las afirmaciones sobre el comportamiento de todo el país puede descomponerse en una serie de pasos:

1. La reunión de los datos en bruto del número de parejas sexuales declaradas por los participantes en nuestra encuesta nos dice algo sobre…

2. El número real de parejas de los miembros de nuestra muestra, que nos dice algo sobre…

3. El número de parejas de los miembros de la población objeto de

estudio —aquellos que potencialmente podrían haber sido incluidos en nuestra encuesta—, que nos dice algo sobre…

4. El número de parejas sexuales de las personas que viven en Gran Bretaña, que es nuestra población objetivo.

¿Cuáles son los eslabones débiles en este razonamiento? Ir de los datos en bruto (estadio 1) a la verdad sobre nuestra muestra (estadio 2) supone hacer algunas asunciones bastante fuertes sobre lo precisos que son los entrevistados cuando declaran cuántas parejas han tenido, y, en realidad, hay muchas razones para la duda. Ya hemos visto que hay una aparente tendencia de los hombres a exagerar, y de las mujeres a minimizar, el número de parejas que han tenido, posiblemente debido a que las mujeres no incluyen a parejas que preferirían olvidar, a diferentes tendencias al redondeo al alza o a la baja, a la mala memoria y al puro y simple «sesgo de aceptabilidad social».[48]

Pasar de nuestra muestra (estadio 2) a la población objeto de estudio (estadio 3) es quizás el paso más difícil. En primer lugar, tenemos que estar seguros de que las personas que han tomado parte en la encuesta son una muestra aleatoria de todas aquellas que resultaban elegibles: esto debería bastar para un estudio bien organizado como Natsal. Pero también tenemos que asumir que las personas que aceptaron formar parte de la encuesta son representativas, y esto es algo menos sencillo. Las encuestas tienen una tasa de respuesta de alrededor de un 66 %, una cifra bastante buena, dada la naturaleza de las preguntas. No obstante, hay cierta evidencia de que las tasas de participación son ligeramente menores entre aquellos que no son tan activos sexualmente, aunque probablemente esto será contrarrestado por la dificultad de conseguir entrevistas con miembros de la sociedad algo menos convencionales.

Finalmente, pasar de la población objeto de estudio (estadio 3) a la población objetivo (estadio 4) es más sencillo, siempre que asumamos que aquellos que potencialmente podrían haber participado representan la población adulta de Gran Bretaña. En el caso de Natsal esto estaría asegurado, dado su sólido diseño experimental, basado en una muestra aleatoria de hogares (lo que no significa que no se haya incluido a personas en instituciones como las cárceles, las fuerzas armadas o los conventos).

Una vez repasadas todas las cosas que pueden ir mal, tal vez haya un cierto escepticismo sobre toda afirmación acerca del comportamiento sexual del país basada en lo que nos dicen los entrevistados en la

encuesta. Pero el objetivo de la ciencia estadística es pasar de un estadio a otro de la mejor manera y finalmente, con toda la humildad de mundo, poder decir qué se puede y qué no se puede aprender de los datos.

Aprendiendo de los datos:

el proceso de «inferencia inductiva»

En los capítulos anteriores hemos visto cómo, ante un problema, se buscan datos, se les echa un vistazo y después se resumen de forma concisa. En ocasiones el recuento, la medición y la descripción son fines en sí mismos. Por ejemplo, si lo único que queremos saber es cuántas personas pasaron por urgencias el año pasado, los datos nos pueden dar la respuesta.

Pero a menudo la pregunta va más allá de la simple descripción de los datos: queremos aprender algo más relevante que las meras observaciones que tenemos delante de nosotros, ya sea para hacer predicciones (¿cuántas personas se pasarán por urgencias el año que viene?) o para decir algo más básico (¿por qué las cifras aumentan?).

Una vez que queremos empezar a generalizar a partir de los datos —aprender algo sobre el mundo más allá de nuestra mera observación— tenemos que preguntarnos sobre qué queremos aprender. Y esto requiere que nos enfrentemos a la inferencia inductiva.

Mucha gente tiene una cierta idea de lo que es la deducción, gracias a Sherlock Holmes y sus razonamientos deductivos al anunciar fríamente que un sospechoso había cometido un delito. En la vida real la deducción es el proceso de usar las reglas frías de la lógica para pasar de las premisas generales a las conclusiones particulares. Si la norma del país es que los coches deben conducir por la derecha, entonces podemos deducir que en cualquier ocasión es mejor conducir por la derecha. Pero la inducción funciona al revés, empezando por casos particulares e intentando llegar de ahí a conclusiones generales. Por ejemplo, supongamos que no sabemos las costumbres de una comunidad sobre cómo besar a las amigas en la mejilla, y tenemos que enterarnos por medio de la observación si la gente se besa una, dos, tres veces o ninguna. La distinción crucial es que la deducción es lógicamente cierta, mientras que la inducción es generalmente incierta.

La figura 3.1 representa la inferencia inductiva como un diagrama genérico, y muestra los pasos que hay que seguir para pasar de los

datos al objetivo final de nuestra investigación: como hemos visto, los datos de la encuesta sobre sexo nos informan del comportamiento de nuestra muestra, que empleamos para obtener a su vez información sobre la gente que podría haber participado en la encuesta, de lo cual podemos extraer algunas conclusiones provisionales sobre el comportamiento sexual de todo el país.

Por supuesto, lo ideal sería poder ir directamente desde los datos en bruto a las afirmaciones generales sobre la población objetivo. En los cursos de Estadística, se asume que las observaciones son completamente aleatorias y proceden directamente de la población de interés. Pero esto es infrecuente en la vida real, y, por tanto, necesitamos considerar todo el proceso que va desde los datos en bruto a nuestro objetivo final. Y, como hemos visto en la encuesta sobre comportamiento sexual, pueden darse problemas en cada uno de los distintos estadios.

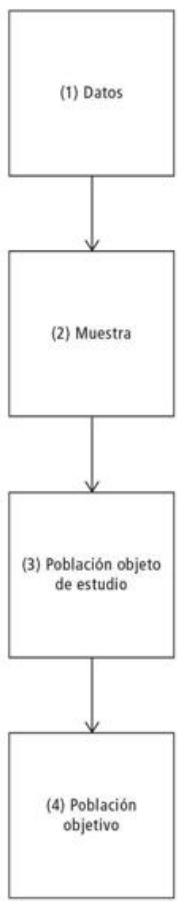


Figura 3.1. Proceso de inferencia inductiva: cada flecha puede ser

interpretada como «nos dicen algo sobre». (Esta estructura en cuatro pasos está tomada de Wayne Oldford).

De los datos (estadio 1) a la muestra (estadio 2): en este caso, se pueden dar problemas de medición: ¿son nuestros datos un reflejo preciso de aquello en lo que estamos interesados? Queremos que nuestros datos sean:

→ Fiables, que tengan poca variación de un caso a otro, que cada cifra sea precisa o repetible.

→ Válidos, en el sentido de medir lo que realmente se quiere medir, y no tener un sesgo sistemático.

Por ejemplo, la validez de la encuesta sobre hábitos sexuales depende de que la gente ofrezca las mismas o similares respuestas a las mismas preguntas cada vez que les pregunten, y esto no debe depender del estilo del entrevistador, o de los caprichos del estado de ánimo o la memoria del entrevistado. Esto puede comprobarse en cierta medida haciendo preguntas específicas al comienzo y al final de la entrevista. La calidad de la encuesta también exige que los entrevistados sean honestos cuando hablan de su actividad sexual, y que ni exageren ni subestimen sistemáticamente sus experiencias. Todas estas exigencias son bastante estrictas.

Una encuesta no sería válida si las respuestas estuviesen sesgadas a favor de una respuesta particular. Por ejemplo, en 2017 la compañía de vuelos baratos Ryanair anunció que el 92 % de sus pasajeros declaraban haber tenido una experiencia de vuelo satisfactoria. Resultó que su encuesta de satisfacción solo permitía las respuestas «Excelente, muy buena, buena, razonable, OK».[49]

Hemos visto cómo el impacto de unos números en el espectador puede verse influido por una presentación positiva o negativa de estos, y, en ese mismo sentido, cómo la presentación de una pregunta puede influir en la respuesta. Por ejemplo, una encuesta de 2015 en el Reino Unido preguntaba a los entrevistados si apoyaban o rechazaban «conceder el derecho al voto a las personas de 16 y 17 años» en el referéndum sobre la permanencia en la Unión Europea; el 52 % estaba de acuerdo, y el 41 %, en contra. Por tanto, la mayoría estaba a favor de la propuesta cuando se planteaba en términos de reconocimiento de derechos y de otorgar poder a los jóvenes.

Pero cuando a los mismos entrevistados se les hizo la pregunta (lógicamente idéntica) sobre si apoyaban o se oponían a «reducir la

edad de votación de 18 a 16 años» para el referéndum, el porcentaje que apoyaba la propuesta cayó al 37 %, mientras que un 56 % se oponía. De manera que cuando la cuestión se planteó en términos de una arriesgada liberalización, la propuesta fue rechazada por la mayoría, un vuelco en la opinión provocado por una sencilla reformulación de la pregunta.[50]

Los entrevistados pueden verse también influidos por lo que se les ha preguntado antes, un proceso conocido como «condicionamiento». Las encuestas oficiales sobre bienestar estiman que alrededor del 10 % de los jóvenes en el Reino Unido se sienten solos, pero un cuestionario online de la BBC encontró un resultado mucho mayor del 42 % entre aquellos que decidieron responder. Esta cifra podría haber resultado inflada por dos factores: el que los datos de la «encuesta» voluntaria sean autorreportados, y el hecho de que la pregunta sobre la soledad hubiese estado precedida por una larga serie de cuestiones acerca de si los encuestados se sentían en líneas generales con falta de relaciones, aislados, dejados de lado y así sucesivamente, todo lo cual podría haberlos condicionado para que diesen una respuesta positiva a la pregunta crucial sobre el sentimiento de soledad.[51]

De la muestra (estadio 2) a la población objeto de estudio (estadio 3): esto depende de la calidad fundamental del estudio, también llamada validez interna: ¿la muestra que observamos refleja fielmente lo que pasa en el grupo que estamos estudiando? Es aquí donde nos encontramos con la forma más importante de evitar sesgos: el muestreo aleatorio. Hasta los niños entienden qué significa escoger algo al azar: cerrar los ojos y meter la mano en una bolsa de dulces y ver cuál sacas, o extraer un número de un sombrero y ver quién gana un premio o una sorpresa (o no). Se ha empleado durante milenios como una forma de asegurar decisiones justas e imparciales, mediante el sorteo,[52] y se ha usado para asignar recompensas,[53] administrar loterías y ocupar posiciones de poder, tanto en cargos públicos como en jurados. También para asignar responsabilidades más graves, como elegir a aquellos jóvenes que han de ir a la guerra o quién puede comer en una balsa a la deriva en medio del mar.

George Gallup, que prácticamente inventó la idea de la encuesta de opinión en la década de 1930, se inventó una analogía muy afortunada para referirse al valor del muestreo aleatorio. Dijo que si tienes que cocinar una gran olla de sopa, no es necesario comértela toda para averiguar si necesita más condimento. Te basta con probar una cucharada, siempre que la hayas removido bien. Una prueba literal de esta idea fue el sorteo utilizado en 1969 para el reclutamiento de la guerra de Vietnam; tenía que proporcionar una

lista ordenada de cumpleaños, y, después, serían reclutados en primer lugar para ir a Vietnam aquellos hombres cuyos cumpleaños estuviesen los primeros de la lista, y así hasta el final de esta. En un intento de hacer que el proceso fuese justo, se prepararon 366 cápsulas, cada una de las cuales contenía un determinado cumpleaños, que se suponía que serían extraídas de una caja de manera aleatoria. Pero las cápsulas fueron introducidas en la caja en el orden del mes del cumpleaños, y no fueron debidamente mezcladas. Esto podría no haber sido problemático si los encargados de extraer las cápsulas hubiesen rebuscado en la caja, pero, como muestra un vídeo, las extraían más bien de la parte superior de la caja.[54] El resultado fue realmente malo para aquellos que nacieron a finales de año: 26 de 31 cumpleaños en diciembre resultaron reclutados, frente a solo 14 en enero.

La idea de «remover» bien es crucial: si quieres generalizar de la muestra a la población, necesitas asegurarte de que tu muestra es representativa. Tener masas de datos no ayuda necesariamente a garantizar una buena muestra y puede incluso dar una sensación falsa de seguridad. Por ejemplo, las empresas demoscópicas fallaron miserablemente en las elecciones generales del Reino Unido de 2015, aunque habían utilizado muestras de miles de votantes potenciales. Una investigación posterior culpó al muestreo no representativo, particularmente en el caso de las encuestas telefónicas. No solo porque la mayoría de las llamadas fueron a líneas fijas, sino porque menos del 10 % de aquellos a los que telefonearon contestaron. Es difícil que esto constituya una muestra representativa.

De la población objeto de estudio (estadio 3) a la población objetivo (estadio 4): finalmente, incluso contando con una medición perfecta y una muestra meticulosamente aleatoria, los resultados podrían seguir sin reflejar lo que queremos investigar si somos incapaces de preguntar a aquellas personas en las que estamos particularmente interesados. Queremos que nuestro estudio tenga validez externa.

Un ejemplo extremo tiene lugar si nuestra población objetivo son personas y solo podemos estudiar animales, como cuando estudiamos el efecto de un producto químico en ratones. Un caso menos dramático es aquel en el que se llevan a cabo pruebas clínicas de un nuevo medicamento únicamente en hombres adultos, pero el medicamento se usa en mujeres y niños. Nos gustaría conocer los efectos en todo el mundo, pero esto no se puede lograr solo a través de un análisis estadístico; inevitablemente necesitamos hacer asunciones y ser muy prudentes.

Cuando tenemos todos los datos

Aunque la idea de aprender de los datos normalmente se asocia a las encuestas, en realidad gran parte de los datos que se emplean en la actualidad no se basan en muestras aleatorias, o, de hecho, en ningún tipo de muestra. Datos cotidianos sobre, por ejemplo, compras en Internet o interacciones en redes sociales, o sobre la administración de sistemas educativos o de policía, pueden ser reutilizados para comprender qué pasa en el mundo. En estos casos, tenemos todos los datos. En términos del proceso inductivo mostrado en la figura 3.1 no hay ninguna separación entre los estadios 2 y 3: la «muestra» y la población objeto de estudio son esencialmente lo mismo. Esto elimina cualquier inquietud acerca del tamaño de la muestra, pero aún quedan otros problemas.

Por ejemplo, preguntémonos sobre cuánta criminalidad hay en Gran Bretaña, y sobre la cuestión políticamente sensible de si está aumentando o disminuyendo. Hay dos fuentes principales de datos, una basada en encuestas y otra en datos administrativos. En primer lugar, la Encuesta sobre Criminalidad en Inglaterra y Gales es un ejemplo clásico de muestreo en el que 38.000 personas son entrevistadas cada año sobre su experiencia con la criminalidad. Al igual que la encuesta Natsal sobre hábitos sexuales, pueden surgir problemas cuando se usan las declaraciones de los entrevistados (estadio 1) para extraer conclusiones sobre sus experiencias reales (estadio 2), dado que los entrevistados podrían no contar la verdad —digamos, por ejemplo, sobre un delito relacionado con drogas en el cual podrían haber participado—. Además, tenemos que asumir que la muestra es representativa de la población elegible, tener en cuenta su tamaño limitado (paso del estadio 2 al estadio 3) y finalmente reconocer que el diseño del estudio puede no alcanzar a parte de la población objetivo global, por ejemplo porque nadie menor de 16 años o que viva en una residencia comunitaria es entrevistado (paso del estadio 3 al estadio 4). No obstante, con todas las salvedades, la Encuesta sobre Criminalidad en Inglaterra y Gales es una «estadística nacional reconocida» y empleada para observar tendencias a largo plazo.[55]

La segunda fuente de los datos son los informes de criminalidad registrados por la policía. Estos datos tienen un objetivo administrativo y no es una muestra: dado que cada delito cometido en el país puede ser contado, la «población objeto de estudio» es la misma que la muestra. Por supuesto, seguimos teniendo que asumir que los datos registrados representan fielmente lo que les pasa a las víctimas que han denunciado un delito (del estadio 1 al estadio 2),

pero el mayor problema se produce cuando queremos afirmar que los datos en la población objeto de estudio —la gente que ha denunciado algún delito— representan la población objetivo de todos los delitos cometidos en Inglaterra y Gales. Desgraciadamente, los registros de la policía dejan fuera sistemáticamente aquellos casos que la policía no clasifica como delito o que no han sido denunciados por la víctima; el consumo ilegal de drogas, por ejemplo, o aquellas personas que deciden no denunciar casos de robo o vandalismo para que los precios de la vivienda en su vecindario no caigan. Un ejemplo extremo es el de los delitos sexuales registrados: después de la publicación de un informe que criticaba las prácticas policiales de clasificación de estos delitos, los crímenes contra la libertad sexual registrados por la policía pasaron de 64.000 en 2014 a 121.000 en 2017, es decir, casi se doblaron en tres años.

No es sorprendente que estas dos fuentes de datos nos lleven a distintas conclusiones sobre las tendencias en la criminalidad. Por ejemplo, la Encuesta sobre Criminalidad estimó que el crimen disminuyó un 9 % entre 2016 y 2017, mientras que la policía registró un aumento del 13 % de los delitos. ¿Cuál de las dos fuentes debemos creernos? Los estadísticos tienen más confianza en la encuesta, y, de hecho, las reservas acerca de la fiabilidad de los datos de delitos registrados por la policía hicieron que perdiesen su designación como estadística nacional en 2014.

Cuando tenemos todos los datos, está muy claro cómo producir estadísticas que describan lo que se ha medido. Pero cuando queremos usar los datos para extraer conclusiones más amplias sobre lo que ocurre a nuestro alrededor, entonces la calidad de los datos pasa a ser algo fundamental, y necesitamos ser conscientes del tipo de sesgo sistemático que puede poner en peligro la fiabilidad de cualquier afirmación que hagamos.

Hay muchas páginas web dedicadas a enumerar los sesgos posibles que se pueden producir en la ciencia estadística, desde el sesgo de asignación (diferencias sistemáticas en quién recibe cada uno de los dos tratamientos médicos objeto de comparación), hasta el sesgo de voluntariedad (el hecho de que aquellos que se presentan voluntarios para un estudio son sistemáticamente diferentes de la población en general). Muchos de estos sesgos son de sentido común, aunque en el capítulo 12 veremos algunos errores más sutiles que se cometen en estadística. Pero primero debemos ocuparnos de las formas de describir nuestro objetivo último: la población objetivo.

La «curva con forma de campana»

Una amiga en Estados Unidos acaba de dar a luz a término a un bebé de 6 libras y 7 onzas (2,91 kilos). Le han dicho que su peso está por debajo de la media, y está preocupada. ¿Es el peso inusualmente bajo?

Ya hemos discutido el concepto de distribución de los datos —la pauta que siguen los datos, conocida normalmente como la distribución empírica o muestral—. A continuación debemos ocuparnos del concepto de distribución poblacional —la pauta que sigue todo el grupo de interés—.

Considérese una mujer estadounidense que acaba de dar a luz. Podríamos pensar que su bebé ha sido extraído, como si fuera una muestra de una única persona, de toda la población de bebés recién nacidos de madres blancas no hispanas en Estados Unidos (su raza es importante, dado que el registro del peso de los recién nacidos se realiza por razas). La distribución de la población es la pauta que asumen los pesos al nacer de todos estos bebés, que podemos obtener del informe del Sistema Nacional de Estados Unidos de Estadísticas Vitales sobre los pesos de alrededor de un millón de bebés nacidos a término en Estados Unidos en 2013 de mujeres blancas no hispanas —aunque este no es un conjunto completo de nacimientos contemporáneos, es una muestra tan grande que podemos considerarla como la población—.[56] Estos pesos al nacer son solo registrados como números en grupos de 500 gramos, y se muestran en la figura 3.2(a).

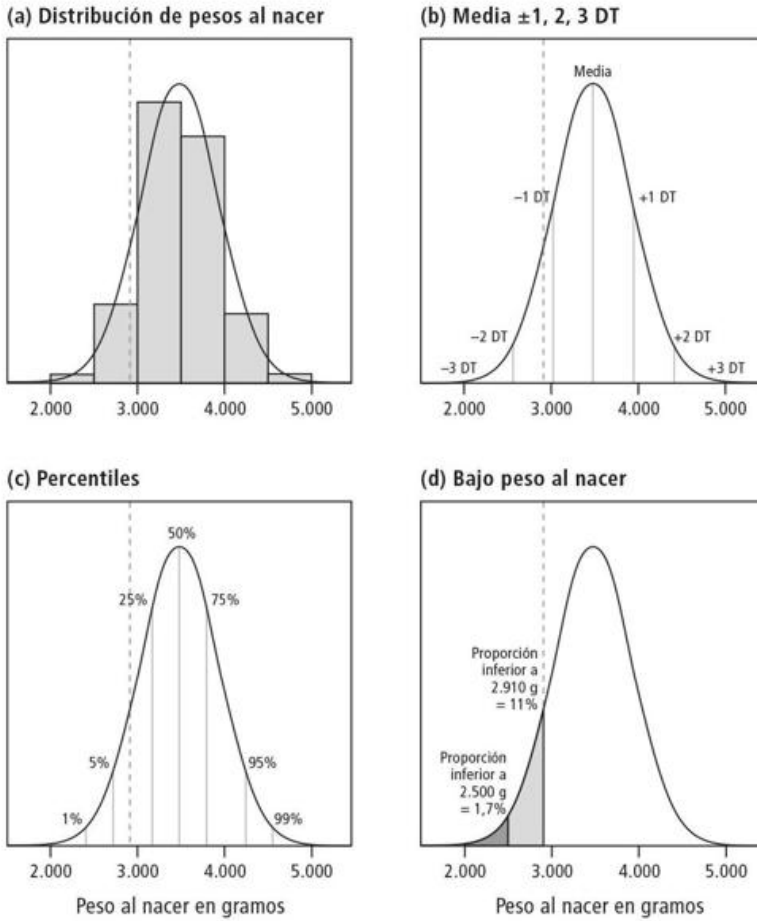


Figura 3.2. (a) Distribución del peso al nacer de 1.096.277 niños de mujeres blancas no hispanas en Estados Unidos en 2013, nacidos en la semana 39-40 de gestación, con una curva normal con la misma media y desviación típica que los pesos registrados de la población. En la línea discontinua se muestra un bebé que pesase 2.910 gramos. (b) Media ± 1, 2, 3 desviaciones típicas (DT) de la curva normal. (c) Percentiles de la curva normal. (d) Proporción de bebés con bajo peso (área sombreada oscura) y bebés con menos de 2.910 gramos (área sombreada clara).

El peso del bebé de su amiga se indica en la línea de 2.910 gramos, y su posición en la distribución puede ser empleada para evaluar si su peso es «inusual». La forma de la distribución es importante. Medidas como el peso, los ingresos, la estatura y demás pueden, al menos en principio, ser tan detalladas como se quiera, y por ello se pueden considerar como cantidades «continuas» cuyas distribuciones

poblacionales tienen una forma suave. El ejemplo clásico es la «curva con forma de campana», o distribución normal, analizada en detalle por primera vez por Carl Friedrich Gauss en 1809 en el contexto de su análisis de los errores de medida en la astronomía y en las encuestas. [57]

La teoría muestra que podemos esperar una distribución normal para fenómenos que están determinados por un gran número de pequeñas variables, por ejemplo un rasgo físico complejo que no está influido solo por unos pocos genes. El peso al nacer, cuando se observa para un único grupo étnico y un único período de gestación, podría considerarse un rasgo de ese tipo, y la figura 3.2(a) muestra una curva normal con la misma media y desviación típica que los pesos registrados. La suave curva normal y el histograma son tranquilizadoramente parecidos, y otros rasgos complejos como la estatura y la capacidad cognitiva también se aproximan a distribuciones poblacionales normales. Otros fenómenos menos naturales podrían tener distribuciones poblacionales que son distintivamente no normales, y que a menudo muestran un larga cola derecha: los ingresos son un ejemplo clásico de ello.

La distribución normal se caracteriza por su media o esperanza, y su desviación típica, que, como hemos visto, es una medida de dispersión —la curva que se ajusta mejor en la figura 3.2(a)—, tiene una media de 3.480 gramos (7 libras, 11 onzas) y una desviación típica de 462 gramos (1 libra). Vemos que las medidas usadas para resumir conjuntos de datos en el capítulo 2 pueden ser aplicadas también para describir una población; la diferencia es que términos como la media y la desviación típica se denominan estadísticos cuando describen un conjunto de datos, y parámetros cuando describen una población. El ser capaz de resumir alrededor de 1.000.000 de mediciones (es decir, alrededor de un millón de nacimientos) en solo dos cantidades es un logro impresionante.

Una gran ventaja de asumir la forma normal para una distribución es que se pueden obtener importantes magnitudes a partir de tablas y software. Por ejemplo, la figura 3.2(b) muestra la posición de la media y de 1, 2 y 3 desviaciones típicas a cada lado de la media. A partir de las propiedades matemáticas de la distribución normal, sabemos que aproximadamente el 95 % de la población se encontrará en el intervalo dado por la media ± 2 desviaciones típicas, y el 99,8 % en ± las 3 desviaciones típicas centrales. El bebé de su amiga está alrededor de 1,2 desviaciones típicas por debajo de la media; esto también se conoce como su puntuación Z, que simplemente mide cuántas desviaciones típicas está alejado de la media un determinado

valor.

La media y la desviación típica pueden usarse como descriptivos de resumen para (la mayoría de) las otras distribuciones, pero también pueden resultar útiles otras medidas. La figura 3.2(c) muestra percentiles seleccionados calculados a partir de la curva normal: por ejemplo, el percentil 50 es la mediana, el punto que divide la población por la mitad y que podría decirse que es el peso de un bebé «promedio» —y que es lo mismo que la media en el caso de una distribución simétrica como la curva normal—. El percentil 25 (3.167 g) es el peso por debajo del cual están el 25 % de los bebés —los percentiles 25 y 75 (3.791 g) se conocen como los cuartiles y la distancia entre ellos (624 g), conocida como rango intercuartílico, es una medida de la dispersión de la distribución—. Una vez más, estos son exactamente los mismos estadísticos de resumen que se usaron en el capítulo 2, pero aquí aplicados a poblaciones, no a muestras.

El bebé de su amiga se encuentra en el percentil 11, lo que significa que el 11 % de los bebés nacidos a término de mujeres blancas no hispanas pesarán menos (la figura 3.2(d) muestra este 11 % como un área sombreada gris claro). Los percentiles del peso al nacer tienen su importancia, dado que el peso del bebé de su amiga será supervisado en relación con el crecimiento esperado para bebés en el percentil 11, [58] y una caída en el percentil en que se encuentra el bebé podría ser motivo de preocupación.

Por razones médicas más que estadísticas, los bebés por debajo de 2.500 gramos son considerados como de «bajo peso al nacer», y los que tienen menos de 1.500 gramos, de «peso muy bajo al nacer». La figura 3.2(d) muestra que podríamos esperar que un 1,7 % de los bebés de este grupo tengan un peso bajo al nacer —en realidad el número exacto era 14.170 (1,3 %), muy cerca de la predicción de la curva normal—. Nótese que este grupo particular de nacimientos a término de madres blancas no hispanas tiene una tasa muy pequeña de pesos bajos: la tasa global para todos los nacimientos en Estados Unidos en 2013 fue de un 8 %, mientras que en el caso de las mujeres negras era del 13 %, una diferencia notable entre razas.

Quizás la lección más importante derivada de este ejemplo es que el área sombreada gris oscuro de la figura 3.2(d) juega dos papeles:

1. Representa la proporción de esta población de bebés que tienen un peso bajo al nacer.

2. Es también la probabilidad de que un bebé escogido aleatoriamente

en 2013 pese menos de 2.500 gramos.

De manera que una población puede ser concebida como un grupo físico de individuos, pero también como la distribución de probabilidad de una observación aleatoria. Esta interpretación dual resultará fundamental cuando nos ocupemos de la inferencia estadística más formal.

Por supuesto, en este caso sabemos la forma y los parámetros de la población, y, por tanto, podemos decir algo tanto de las proporciones dentro de la población como de las probabilidades de distintos eventos derivados de una observación aleatoria. Pero lo que este capítulo busca recalcar es que normalmente no conocemos los datos de la población, y por ello queremos seguir un proceso inductivo para ir, al revés, de los datos a la población. Hemos visto que las medidas habituales de media, mediana, moda y así sucesivamente, que hemos desarrollado a partir de muestras, se extienden a todas las poblaciones; la diferencia es que no sabemos cuáles son. Y ese es el desafío al que nos enfrentaremos en el siguiente capítulo.

¿Qué es la población?

Los distintos estadios de la inducción descritos más arriba funcionan bien en el caso de encuestas planificadas, pero muchos análisis estadísticos no encajan completamente en ese marco. Hemos visto que, especialmente cuando se usan registros administrativos tales como informes policiales sobre delincuencia, podríamos disponer de todos los datos posibles. Pero, aunque no haya muestreo, la idea de una población subyacente puede seguir siendo útil.

Considérense los datos de cirugía cardiaca infantil del capítulo 1. Hicimos entonces la asunción bastante osada de que no había problemas de medición —en otras palabras, que tenemos información tanto de las operaciones como de los supervivientes a treinta días en cada hospital—. Por tanto, nuestro conocimiento de la muestra (estadio 2) es perfecto.

Pero ¿cuál es la población objeto de estudio? Tenemos datos sobre todos los niños y todos los hospitales, de manera que no existe un grupo mayor del que hayamos extraído una muestra. Aunque la idea de la población se menciona casi de manera casual en los cursos de estadística, este ejemplo muestra que se trata de una idea difícil y sofisticada que vale la pena explorar con cierto detalle, puesto que muchas ideas importantes se basan en este concepto.

Hay tres tipos de poblaciones de las cuales se podría extraer una muestra, ya se refieran los datos a personas, transacciones, árboles o lo que sea.

→ Una población literal. Se trata en este caso de un grupo identificable, como cuando escogemos una persona de manera aleatoria cuando hacemos encuestas. O podría tratarse de un grupo de individuos que puedan ser medidos, y aunque no escogemos realmente uno de manera aleatoria, tenemos datos de voluntarios. Por ejemplo, podríamos considerar que las personas que predecían el número de gominolas eran una muestra de una población de todos los fanáticos de las matemáticas que ven vídeos de YouTube.

→ Una población virtual. Normalmente tomamos medidas mediante algún tipo de dispositivo, como cuando tomamos la tensión o medimos la polución atmosférica. Sabemos que podríamos tomar nuevas mediciones y obtener resultados ligeramente distintos, como usted sin duda sabrá si se ha tomado la tensión repetidamente. Lo parecidos que sean los resultados obtenidos depende de la precisión del dispositivo utilizado y de la estabilidad de las circunstancias —podemos pensar en ello como si se tratase de extraer observaciones de una población virtual de todas las medidas que podrían ser tomadas si tuviésemos tiempo—.

→ Una población es metafórica si no existe ninguna población mayor. Se trata de un concepto inusual. En este caso, hacemos como si los datos fuesen extraídos aleatoriamente de alguna población, pero claramente no es el caso, como en el ejemplo de los niños sometidos a cirugía cardiaca: no realizamos ningún muestreo, tenemos todos los datos, y no queda ninguno por recopilar. Piénsese en el número de asesinatos que se producen al año, los resultados de los exámenes en una clase concreta o datos de todos los países del mundo: ninguno de ellos se puede considerar una muestra de una población.

La idea de una población metafórica es bastante compleja, por lo que podría ser mejor imaginarnos que lo que hemos observado se ha extraído de un espacio imaginario de posibilidades. Por ejemplo, la historia del mundo es la que es, pero podríamos imaginarnos cómo habría sido si todo hubiese resultado de manera diferente y nosotros hubiésemos acabado en solo uno de esos posibles estados del mundo. Este conjunto de todas las historias alternativas puede ser considerado una población metafórica. Por ser más concretos, cuando observábamos los datos de la cirugía cardiaca infantil en el Reino Unido entre 2012 y 2015, teníamos todos los datos para todos esos años, y sabíamos cuántas muertes se produjeron y cuántos niños

sobrevivieron. Aun así, podríamos imaginar historias contrafactuales en las cuales habrían sobrevivido individuos distintos, por circunstancias imprevistas que tendemos a denominar «azar».

A estas alturas debería haber quedado claro que pocas aplicaciones de la ciencia estadística implican un muestreo aleatorio literal, y que es cada vez más común tener todos los datos potencialmente disponibles. No obstante, es muy importante comprender la idea de una población imaginaria de la cual se extrae nuestra «muestra», porque de esta manera podemos usar todas las técnicas matemáticas que hemos desarrollado para las muestras de poblaciones reales.

Personalmente, me gusta actuar como si todo lo que ocurre a nuestro alrededor fuese un resultado aleatorio de entre todos los que podrían darse. Depende de nosotros creer que es algo realmente aleatorio, ya sea debido a la voluntad de un dios o de muchos dioses, o a cualquier otra teoría de causalidad: escojamos una cosa u otra, a las matemáticas les es indiferente. Es solo una forma de estrujarnos la mente para poder aprender mejor de los datos.

Resumen

→ La inferencia inductiva requiere empezar desde los datos, y pasar de ahí a la muestra, a la población objeto de estudio y, finalmente, a la población objetivo.

→ A todo lo largo de este camino pueden surgir problemas y sesgos.

→ La mejor manera de pasar de la muestra a la población objeto de estudio es mediante una muestra aleatoria.

→ Una población puede ser un grupo de individuos, pero también la distribución de probabilidad de una observación aleatoria extraída de esa población.

→ Las poblaciones pueden ser sintetizadas con parámetros similares a los estadísticos de resumen de los datos de muestras.

→ A menudo los datos no son una muestra de una población literal. Cuando tenemos todos los datos existentes, entonces podemos imaginarnos que han sido extraídos de una población metafórica de eventos que podrían haber ocurrido, pero no lo hicieron.

[48] Un experimento aleatorizado con estudiantes estadounidenses

proporcionó cierta evidencia de ese sesgo. Mujeres conectadas a un detector de mentiras tendían a admitir haber tenido más parejas que aquellas a las que se garantizaba el anonimato, mientras que ese efecto no se encontró en los hombres. No se dijo a los participantes que el detector de mentiras era falso.

[49] Después de que alguien de la Real Sociedad Estadística criticase la encuesta, un portavoz del jefe de Ryanair, Michael O'Leary, dijo que «el 95 % de los clientes de Ryanair no han oído hablar de la Real Sociedad Estadística, al 97 % no le importa lo que diga y el 100 % piensa que da la impresión de que sus miembros necesitan reservar una oferta vacacional de Ryanair». En otra encuesta de esa misma época, Ryanair fue valorada como la peor de veinte aerolíneas europeas (aunque también aquí había problemas de fiabilidad, dado que la encuesta se llevó a cabo cuando Ryanair había cancelado un gran número de vuelos).

[50] Ipsos MORI, ¿Qué piensa el Reino Unido?, 2015, https://whatukthinks.org/eu/poll/ipsos-mori-141215.

[51] Datos del programa More or Less, 5 de octubre de 2018, https://www.bbc.co.uk/programmes/p06n2lmp. La demostración clásica de los efectos del condicionamiento es la comedia británica Sí, primer ministro, cuando el alto funcionario sir Humphrey Appleby muestra cómo las preguntas adecuadas pueden proporcionar las respuestas requeridas. Este ejemplo se usa ahora en la enseñanza de métodos estadísticos, https://researchmethodsdataanalysis.blogspot.com/2014/01/leading-questions-yes-prime-minister.html.

[52] No confundir con el sortilegio, una forma de adivinación en la cual fenómenos aparentemente aleatorios son empleados para determinar la voluntad divina o la fortuna futura —esto también se conoce como cleromancia—. Existen ejemplos de ello en muchas culturas, por ejemplo la adivinación mediante hojas del té o vísceras de pollo, los ejemplos bíblicos en los que se echan suertes para determinar la voluntad de Dios y la adivinación usando el I Ching.

[53] «Y Jesús decía: Padre, perdónalos, porque no saben lo que hacen. Y repartieron entre sí sus vestidos, echando suertes», Lucas 23, 34.

[54] El vídeo está en https://www.youtube.com/watch?v=-p5X1FjyD_g. Véase también http://www.historynet.com/whats-your-number.htm.

[55] Detalles sobre la Encuesta sobre Criminalidad en Inglaterra y Gales y sobre registros de criminalidad de la policía pueden obtenerse en la Oficina de Estadísticas Nacionales: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice.

[56] Los pesos de los nacimientos en Estados Unidos están en: http://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_01.pdf.

[57] Lo que descubrió Gauss no se basó en la observación empírica, sino que era una forma teórica de error de medición que justificase sus métodos estadísticos.

[58] Aunque las distribuciones usadas para esta supervisión serán algo más sofisticadas que la normal.

04

¿Qué causa qué?

¿Ir a la universidad incrementa el riesgo de desarrollar un tumor cerebral?

La epidemiología es el estudio de por qué se producen enfermedades en la población, y los países escandinavos son el sueño de todo epidemiólogo. Esto se debe a que en esos países todo el mundo tiene un número de identificación personal que se usa para registrarse en los servicios de sanidad, educación, impuestos y demás, y esto permite que los investigadores relacionen estos distintos aspectos de la vida de las personas de una manera que sería imposible (y quizás políticamente controvertida) en otros países.

Se realizó un estudio particularmente ambicioso que implicaba a unos cuatro millones de hombres y mujeres suecos cuyos datos fiscales y sanitarios fueron relacionados a lo largo de dieciocho años, lo que permitió a los investigadores concluir que los hombres con una mejor posición socioeconómica tenían una tasa ligeramente mayor de ser diagnosticados con un tumor cerebral. Se trataba de uno de esos estudios valiosos pero no muy excitantes que normalmente no atraerían mucha atención, por lo que un empleado de la oficina de divulgación de la universidad pensó que resultaría más interesante decir en la nota de prensa correspondiente que «los niveles altos de educación están vinculados a un incremento del riesgo de tumor cerebral», aunque el estudio era sobre la posición socioeconómica más que sobre la educación. Y para cuando esta información hubo alcanzado a la opinión pública, a un subeditor de un periódico se le ocurrió el titular de turno: «Por qué ir a la universidad incrementa el riesgo de desarrollar un tumor cerebral».[59]

Para cualquiera que tenga años de credenciales académicas, este titular de prensa podría resultar alarmante. Pero ¿deberíamos preocuparnos? Se trataba de un estudio gigantesco basado en un registro de toda la población elegible —no de una muestra—, de manera que podíamos concluir con bastante seguridad que las personas con mayor nivel educativo desarrollan un número ligeramente mayor de tumores cerebrales. Pero ¿realmente todo el sudor acumulado en las bibliotecas acaba recalentando el cerebro y esto lleva a extrañas mutaciones celulares? A pesar del titular anteriormente mencionado, lo dudo. Y hay que reconocer en su honor que los autores del artículo también lo dudaban, afirmando que «la

exhaustividad de los registros del cáncer y el sesgo de detección son explicaciones potenciales de los resultados». En otras palabras, es más probable que las personas más ricas con educación superior sean diagnosticadas y que su tumor acabe siendo registrado, un ejemplo de lo que se conoce como la falacia de la evidencia incompleta en epidemiología.

«La correlación no implica causalidad»

Hemos visto en el anterior capítulo cómo el coeficiente de correlación de Pearson mide la proximidad de los puntos de un diagrama de dispersión a una línea recta. En el caso de los hospitales ingleses que llevaban a cabo cirugía cardiaca infantil en la década de 1990, y cruzando el número de casos con las tasas de supervivencia, la alta correlación indicaba que los mayores hospitales estaban asociados con una menor mortalidad. Pero a partir de ahí no podríamos concluir que los mayores hospitales causaron esa menor mortalidad.

Esta actitud cautelosa tiene un largo pedigrí. Cuando el recién creado coeficiente de correlación de Pearson estaba siendo debatido en la revista Nature en 1900, un comentarista advirtió que «la correlación no implica causalidad». En el siglo siguiente esta frase se ha convertido en un mantra repetido una y otra vez por los estadísticos cuando se enfrentan a afirmaciones basadas simplemente en la observación de que dos cosas tienden a variar juntas. Hay incluso una página web que genera automáticamente asociaciones absurdas, como, por ejemplo, una especialmente deliciosa que muestra una correlación de 0,96 entre el consumo anual per cápita de queso mozzarella en Estados Unidos entre 2000 y 2009 y el número de doctorados en Ingeniería Civil en cada uno de esos años.[60]

Parece existir una necesidad profundamente arraigada en los humanos de explicar las cosas que pasan en términos de una relación simple causa-efecto —estoy seguro de que todos podemos inventarnos una buena historia sobre todos esos nuevos ingenieros hinchándose de pizzas—. Hay incluso una palabra para la tendencia a justificar una conexión entre eventos que no están relacionados, apofenia, y el caso más extremo es aquel en el que se culpa de una desgracia o de simple mala suerte a la mala voluntad de otros o incluso a la brujería.

Por desgracia, o quizás afortunadamente, el mundo es un poco más complicado que atribuir todo a la brujería. Y la primera complicación viene de intentar averiguar lo que queremos decir cuando hablamos de «causa».

Pero, en realidad, ¿qué es la «causalidad»?

La causalidad es una cuestión muy controvertida, lo cual es quizá un tanto sorprendente, dado que en la vida real parece algo bastante sencillo: hacemos algo, y esto lleva a otra cosa. Me pillé el pulgar con la puerta del coche, y ahora me duele.

Pero ¿cómo sabemos que mi pulgar no me habría dolido de todas maneras? Quizás para ello podemos pensar en lo que se conoce como contrafactual. Si no me hubiese pillado el pulgar con la puerta, entonces mi pulgar no me dolería. Pero esto siempre será un supuesto, que requiere reescribir la historia, dado que nunca sabremos con certeza qué es lo que podría haber sentido (aunque en este caso estoy bastante seguro de que mi pulgar no me habría empezado a doler por su cuenta).

La cosa se pone aún más interesante cuando incluimos la inevitable variabilidad que acompaña a todo lo bueno que tiene la vida. Por ejemplo, hay ahora consenso entre la comunidad médica sobre que fumar produce cáncer, pero han pasado décadas hasta que los médicos han alcanzado esta conclusión. ¿Por qué les llevó tanto tiempo? Porque la mayoría de los fumadores no desarrollan cáncer de pulmón, y algunas personas que no fuman aun así lo desarrollan. Todo lo que se puede decir es que es más probable tener cáncer de pulmón si se fuma que si no, y esta es una de las razones por las cuales se ha tardado tanto en aprobar leyes que restrinjan el consumo de tabaco.

De manera que nuestra idea «estadística» de la causalidad no es estrictamente determinista. Cuando decimos que X causa Y, no queremos decir que cada vez que X se produce, Y también lo hace. O que Y solo ocurrirá si ocurre X. Simplemente queremos decir que si intervenimos y forzamos a que ocurra X, entonces Y tiende a ocurrir con más frecuencia. De manera que nunca podemos decir que X causó Y en un caso específico, solo que X incrementa el porcentaje de veces en que pasa Y. Esto tiene dos consecuencias fundamentales que afectan a qué es lo que tenemos que hacer si queremos saber qué causa qué. En primer lugar, para inferir causalidad con total seguridad, idealmente necesitamos intervenir y llevar a cabo experimentos. En segundo lugar, dado que este es un mundo estadístico o estocástico, necesitamos intervenir más de una vez para reunir evidencia.

Y esto nos lleva a un tema delicado: llevar a cabo experimentos médicos sobre un gran número de personas. A pocos de nosotros nos atrae la idea de ser el objeto de experimentos, especialmente si es una

cuestión de vida o muerte. Esto hace que sea aún más sorprendente que miles de personas hayan estado dispuestas a formar parte de enormes estudios en los que ni ellos ni sus médicos sabían qué tratamiento acabarían recibiendo.

¿Reducen las estatinas los infartos y los ictus?

Todos los días me tomo una pastillita blanca, una estatina, porque me han dicho que reduce el colesterol y, de esa manera, el riesgo de infartos e ictus. Pero ¿cómo me afecta a mí personalmente? Estoy seguro de que hace que mi colesterol de baja densidad (LDL, por sus siglas en inglés) disminuya, dado que me dijeron que se había reducido al poco de comenzar a tomar las pastillas. Esta caída del LDL es un efecto directo, esencialmente determinista, que puedo asumir que ha sido causado por la estatina.

Pero nunca sabré si esta costumbre diaria es buena para mí a largo plazo: depende de cuál de mis muchas vidas futuras posibles se da realmente. Si nunca tengo un infarto o un ictus, no tendré ni idea de si nunca hubiera tenido uno incluso aunque no hubiese tomado las pastillas, en cuyo caso este consumo de pastillas años y años habría sido una pérdida de tiempo. Si tengo un infarto o un ictus, no sabré si ese evento se retrasó gracias a la estatina. Todo lo que podré llegar a saber es que de media beneficia a un gran número de personas como yo, y que este conocimiento se basa en pruebas clínicas.

El propósito de una prueba clínica es llevar a cabo un «test justo» que determine adecuadamente la causalidad y estime el efecto medio de un nuevo tratamiento médico, sin introducir sesgos que nos puedan dar la idea equivocada de su efectividad.

Una prueba médica adecuada debería idealmente obedecer a los siguientes principios:

1. Controles. Si queremos investigar el efecto de las estatinas en una población, no podemos simplemente dar estatinas a unas pocas personas y, después, si no tienen un infarto, afirmar que ha sido consecuencia de la píldora (con independencia de las páginas web que usan esta forma de razonamiento anecdótico para vender sus productos). Necesitamos un grupo de intervención, al que se darán las estatinas, y un grupo de control, al que se darán pastillas de azúcar o placebos.

2. Asignación del tratamiento. Es importante comparar cosas parecidas, por lo que los grupos de tratamiento y de comparación

tienen que ser tan similares como sea posible. La mejor manera de asegurar esto es asignando a los participantes aleatoriamente a ser tratados o no, y después ver qué les ocurre —esto se conoce como un ensayo aleatorio controlado (EAC)—. Las pruebas con estatinas emplean a suficientes personas para que los dos grupos sean similares en todos los factores que podrían influir en el resultado, incluyendo —y esto es especialmente importante— aquellos factores que desconocemos. Estos estudios pueden ser enormes: en el Estudio del Reino Unido para la Protección Coronaria llevado a cabo a finales de la década de 1990, 20.536 personas con un alto riesgo de infarto o ictus fueron aleatoriamente asignadas a tomar o bien 40 miligramos de simvastatina diariamente o bien una pastilla falsa.[61]

3. Las personas deben ser medidas en los grupos a los que han sido asignadas. Las personas asignadas al grupo de la «estatina» en el Estudio de Protección Cardiaca (HPS)[62] fueron incluidas en el análisis final incluso aunque no se hubiesen tomado sus estatinas. Esto se conoce como el principio de intención de tratar, y puede parecer bastante extraño. Significa que la estimación final del efecto de las estatinas mide realmente el efecto de haber sido prescrito con estatinas, y no tanto de haberlas tomado. En la práctica, por supuesto, se animará a los participantes a que se tomen las pastillas a lo largo del estudio, aunque después de cinco años en el HPS, el 18 % de aquellos a los que se asignó el consumo de estatina dejaron de tomarla, mientras que hasta un 32 % de aquellos asignados inicialmente a un placebo de hecho empezaron a tomar estatinas durante el estudio. Dado que estas personas que cambian de tratamiento tienden a embarrar las diferencias entre los grupos, podríamos esperar que el efecto aparente en un análisis de «intención de tratar» fuese menor que el efecto de tomar realmente el medicamento.

4. Si es posible, los participantes no deberían ni siquiera saber en qué grupo están. En el estudio de las estatinas, tanto las estatinas de verdad como el placebo parecen idénticos, de manera que los participantes eran ciegos al tratamiento que estaban recibiendo.

5. Los grupos deben ser tratados por igual. Si el grupo asignado a la estatina fuese invitado a citas hospitalarias más frecuentes, o examinado más atentamente, sería imposible separar los beneficios del medicamento de los beneficios de unos mayores cuidados médicos. En el HPS, el personal que llevaba a cabo el seguimiento clínico no sabía si los pacientes estaban tomando una estatina o un placebo, por lo que ellos también estaban ciegos a la asignación del tratamiento.

6. Si es posible, aquellos que evalúan los resultados finales no deben saber en qué grupo están los sujetos. Si un médico cree que un tratamiento funciona, podría, debido a un sesgo inconsciente, exagerar el beneficio del grupo de tratamiento.

7. Medir a todo el mundo. Hay que hacer todos los esfuerzos necesarios para seguir a todos los participantes, dado que aquellos que abandonan el estudio podrían, por ejemplo, haberlo hecho por los efectos secundarios del medicamento. El HPS tiene un estimable 99,6 % de seguimiento completo a cinco años, con los resultados que se muestran en la tabla 4.1.

Aquellos que fueron asignados al grupo de la estatina claramente tuvieron de media mejores resultados de salud, y dado que los pacientes habían sido aleatorizados y tratados, por lo demás, de manera idéntica, puede asumirse que este resultado es un efecto causal debido a habérseles prescrito estatinas. Pero hemos visto que muchas personas no se atuvieron realmente al tratamiento al que habían sido asignadas, y esto lleva a que en cierta medida se diluya la diferencia entre los grupos: los investigadores del HPS estiman que el efecto real de tomar estatinas es alrededor de un 50 % mayor que el que se muestra en la tabla 4.1.

Evento	Porcentaje de 10.267 personas asignadas al placebo	Porcentaje de las 10.269 personas asignadas a la estatina	% de riesgo relativo de reducción entre aquellos asignados a las estatinas
Infarto	11,8	8,7	27%
Ictus	5,7	4,3	25%
Muerte por cualquier causa	14,7	12,9	13%

Tabla 4.1. Resultados a cinco años en el Estudio de Protección Cardiaca, de acuerdo con tratamientos asignados a pacientes. La reducción absoluta del riesgo de infarto fue de 11,8 – 8,7 = 3,1 %. De manera que, de cada 1.000 personas que tomaron estatina, se evitaron alrededor de 31 infartos; esto significa que alrededor de 30 personas tuvieron que tomar una estatina durante 5 años para prevenir un infarto.

Dos puntos finales muy importantes:

8. No hay que basarse en un único estudio. Un único estudio sobre la estatina nos muestra que el medicamento ha funcionado en un caso particular, pero unas conclusiones sólidas requieren múltiples estudios.

9. La evidencia ha de ser revisada sistemáticamente. Cuando se examinan múltiples estudios, hay que asegurarse de incluir cada uno de los estudios realizados, y, de esta manera, crear lo que se conoce como una «revisión sistemática». Los resultados podrían ser combinados formalmente en un metaanálisis.

Por ejemplo, una reciente revisión sistemática reunió evidencia de 27 análisis aleatorizados de estatinas, que incluían a más de 170.000 personas con bajo riesgo de enfermedad cardiovascular.[63] Pero en lugar de centrarse en la diferencia entre los grupos asignados a tomar estatinas y controles, se centraban en estimar el efecto en la reducción del LDL. Asumían esencialmente que el efecto de la estatina consiste en cambiar los lípidos en la sangre, y basaron sus cálculos en la reducción media del LDL apreciada en cada estudio, teniendo en cuenta cualquier incumplimiento del tratamiento asignado. Con esta asunción extra sobre los mecanismos a través de los cuales las estatinas benefician a nuestra salud, pudieron estimar el efecto de tomarse la estatina, que, concluyeron, era una reducción del 21 % de grandes eventos vasculares por cada reducción de 1 mmol/l (milimoles por litro) de colesterol LDL. Que es bastante para que siga tomando mis pastillas.[64]

Hemos ignorado la posibilidad de que una relación observada no sea causal, sino simplemente el resultado del puro azar. La mayoría de los medicamentos en el mercado tienen solo unos efectos moderados, solo ayudan a una minoría de las personas que los toman y su beneficio global solo puede ser detectado con seguridad a través de grandes y meticulosos estudios aleatorizados. Los estudios sobre la estatina son enormes, especialmente cuando se monta un metaanálisis, lo que significa que los resultados que hemos discutido no pueden atribuirse al azar (veremos cómo comprobar esto en el capítulo 10).

¿Es efectivo rezar?

La lista de los principios para las Pruebas Controladas Aleatorizadas (PCA) no es nueva: casi todos ellos fueron introducidos en 1948 en lo que normalmente se considera el primer ensayo clínico propiamente dicho. Se analizaban los efectos de la estreptomicina, un medicamento prescrito para combatir la tuberculosis. Era muy osado asignar a unos pacientes a este tratamiento que potencialmente podía salvar sus vidas y a otros no, pero la decisión se vio favorecida por el hecho de que no había suficientes dosis del medicamento para todo el mundo en esos momentos en el Reino Unido, por lo que una asignación aleatoria parecía una forma justa y ética de decidir quién obtendría el tratamiento. Y, no obstante, mucho tiempo después, y tras miles de

análisis PCA, el público todavía se sorprende de que unas decisiones médicas acerca de qué tratamiento recomendar a un individuo, incluso algunas tan dramáticas como si someterse a una mastectomía radical o una tumorectomía para combatir el cáncer de mama, se resuelvan tirando una moneda al aire (aunque se trate de una moneda metafórica, materializada en un generador computarizado de números aleatorios).[65]

En la práctica, el proceso de asignar tratamientos en este tipo de estudios es generalmente más complejo que la mera aleatorización caso por caso, dado que queremos asegurarnos de que todos los tipos de personas están igualmente representados en los grupos a los que se asigna distintos tratamientos. Por ejemplo, podríamos querer tener aproximadamente el mismo número de personas mayores con alto riesgo en el grupo de las estatinas y de los placebos. Esta idea proviene de los experimentos agrícolas, donde se originaron muchas de las ideas de los estudios aleatorizados, en gran medida a partir del trabajo de Ronald Fisher (del que hablaremos más adelante). Por ejemplo, un gran campo podría dividirse en terrenos individuales, y después se asignaría aleatoriamente un fertilizante diferente a cada una de estas explotaciones, de la misma manera que las personas son asignadas aleatoriamente a un tratamiento médico. Pero el campo podría contener partes que fuesen sistemáticamente diferentes debido al drenaje, lo sombreadas que sean y así sucesivamente, y, por ello, el campo debe ser primero dividido en «bloques» que contengan terrenos que sean aproximadamente similares. La aleatorización debe organizarse entonces de tal manera que quede garantizado que cada bloque contenga un número igual de terrenos dado cada fertilizante, lo que significaría, por ejemplo, que los tratamientos estarían equilibrados en lo que respecta a las áreas pantanosas.

Por ejemplo, en una ocasión yo trabajé en un estudio aleatorizado que comparaba métodos alternativos para curar hernias: la cirugía abierta estándar versus la cirugía laparoscópica o cirugía mínimamente invasiva. Se sospechaba que la pericia del equipo podría incrementarse durante el estudio, y por ello era esencial que los dos tratamientos estuvieran equilibrados en todo momento, a medida que el estudio avanzaba. Por ello agrupé la secuencia de pacientes en bloques formados por grupos de 4 y 6, y me aseguré de que los pacientes fueran aleatorizados de la misma manera a cada tratamiento dentro de cada bloque. En esa época, los tratamientos estaban escritos en trocitos de papel, que doblaba y metía en sobres opacos de color marrón numerados. Recuerdo ver a los pacientes tumbados en la camilla preoperatoria, sin tener ni idea de qué tratamiento les iba a tocar, mientras el anestesista abría el sobre para revelar lo que les iba

a pasar, y, más en concreto, si se iban a ir a casa con una gran cicatriz o con una serie de pinchazos.

Los estudios aleatorizados se convirtieron en la regla de oro para probar nuevos tratamientos médicos, y ahora se usan crecientemente para estimar los efectos de las nuevas políticas educativas y policiales. Por ejemplo, el Equipo de Análisis del Comportamiento del Reino Unido solicitó aleatoriamente a la mitad de los estudiantes que volvían a examinarse del certificado general de educación secundaria en matemáticas o inglés que seleccionasen a alguien que les enviase mensajes de texto regulares animándolos en sus estudios; los estudiantes con un «promotor de sus estudios» tuvieron una tasa de aprobados un 27 % mayor. El mismo equipo también observó una variedad de efectos positivos en un estudio aleatorizado de cámaras portátiles para oficiales de policía, como por ejemplo que un número menor de personas eran innecesariamente detenidas y registradas.[66]

Siempre ha habido estudios para determinar la efectividad de las plegarias. Por ejemplo, el Estudio de los Efectos Terapéuticos de la Oración Intercesora (STEP)[67] asignó aleatoriamente a alrededor de 1.800 pacientes cardiacos con un baipás a tres grupos: los pacientes en los grupos 1 y 2 eran objeto de oraciones o no, respectivamente, pero no sabían si era así o no, mientras que los del grupo 3 sabían que estaban rezando por ellos. El único efecto aparente fue un pequeño incremento en las complicaciones en el grupo que sabía que estaban rezando por ellos: uno de los investigadores comentó: «Les debió dejar preocupados, preguntándose: ¿estoy tan enfermo que han tenido que llamar a su equipo de plegarias?».[68]

La principal innovación reciente en experimentación aleatorizada se refiere a la prueba «A/B» en diseño de páginas web, en la cual los usuarios son (sin saberlo) dirigidos a diseños alternativos de páginas web, y se mide el tiempo invertido en cada página, los clics a anuncios, *etc.* Una serie de pruebas A/B pueden llevar rápidamente a un diseño optimizado, y los enormes tamaños muestrales significan que se pueden detectar incluso los efectos más pequeños, pero no por ello no lucrativos. Esto significa que toda una nueva comunidad de profesionales ha tenido que aprender sobre diseño de estudios experimentales, incluidos los peligros de las comparaciones múltiples, algo a lo que nos referiremos en el capítulo 10.

¿Qué hacemos cuando no podemos aleatorizar?

¿Por qué los hombres ancianos tienen las orejas grandes?

Para los investigadores, es fácil aleatorizar si todo lo que tienen que hacer es cambiar una página web: no hay que realizar ningún esfuerzo para contratar participantes dado que ni siquiera saben que son los sujetos de un experimento, y no hay necesidad de obtener aprobación ética para usarlos como conejillos de Indias. Pero la aleatorización es en ocasiones difícil y a veces imposible: no podemos comprobar el efecto de nuestros hábitos asignando aleatoriamente a las personas a tratamientos en los que tienen que fumar o seguir una dieta poco saludable (aunque esos experimentos sí se realizan con animales). Cuando los datos no provienen de un experimento, se dice que son observacionales. Por tanto, a menudo tenemos que hacer lo que podamos para separar la correlación de la causalidad utilizando un buen diseño y técnicas estadísticas aplicadas a datos observacionales, combinado todo ello con una saludable dosis de escepticismo.

Puede que el tema de las orejas de los hombres mayores sea mucho menos importante que algunos de los otros temas del libro, pero ilustra la necesidad de escoger diseños de investigación apropiados para responder las preguntas. Si adoptamos un enfoque de resolución de problemas basado en el ciclo PPDAC, el Problema consiste en que (basándome en mi observación personal), los hombres ancianos parecen tener a menudo orejas grandes. ¿Por qué es así? Un Plan obvio es ver si, en la población contemporánea, la edad está correlacionada con el tamaño de las orejas de los adultos. Varios equipos de investigadores médicos en el Reino Unido y Japón han reunido datos en un estudio transversal: su Análisis mostró una clara correlación positiva, y su Conclusión fue que el tamaño de las orejas está asociado con la edad.[69]

El desafío es, entonces, intentar explicar esta asociación. ¿Continúan creciendo las orejas con la edad? ¿O bien las personas que son ahora ancianas siempre han tenido orejas grandes, y algo ha ocurrido a lo largo de las últimas décadas que hace que las generaciones más recientes tengan orejas más pequeñas? O podría ser que los hombres con orejas más pequeñas simplemente mueren antes por alguna razón; hay una creencia china tradicional según la cual unas orejas grandes son señal de una vida larga. Necesitamos mucha imaginación para idear qué tipo de estudio podría comprobar estas ideas. Un estudio prospectivo de cohorte seguiría a los hombres jóvenes a lo largo de sus vidas, midiendo sus orejas para comprobar si han crecido, o si aquellos con orejas más pequeñas han muerto antes. Esto llevaría mucho tiempo, por lo que podríamos optar, en su lugar, por un estudio retrospectivo de cohorte, que se centraría en hombres que son ahora ancianos e intentaría determinar si sus orejas habían crecido, quizás usando evidencia fotográfica pasada. Un estudio de caso y

control podría centrarse en hombres que han muerto, encontrar hombres todavía vivos con la misma edad y con otros factores similares que se sepa que predicen la longevidad, y ver si los supervivientes tienen orejas más grandes.[70]

Y, así, el ciclo de resolución de problemas empezaría otra vez.

¿Qué podemos hacer cuando observamos una asociación?

Es aquí donde se requiere una cierta imaginación estadística; puede ser divertido pensar en las razones por las cuales una correlación observada podría ser espuria. Algunas son muy sencillas: la fuerte correlación entre consumo de mozzarella y número de ingenieros civiles se debe probablemente a que ambas magnitudes se han incrementado con el tiempo. Igualmente, cualquier correlación entre las ventas de helados y los ahogamientos se debe a que ambos están influidos por el tiempo atmosférico. Cuando una asociación aparente entre dos resultados podría ser explicada por algún factor común observado que influye en ambos, esta causa común se conoce como variable de confusión: tanto el año como el tiempo atmosférico son variables de confusión potenciales, dado que pueden ser medidas y consideradas en un análisis.

La técnica más sencilla para enfrentarse a una variable de confusión es analizar la relación aparente en cada nivel de esa variable. Esto se conoce como ajuste o estratificación. Así, por ejemplo, podríamos explorar la relación entre ahogamientos y ventas de helados en días con aproximadamente la misma temperatura.

	Mujeres			Hombres		
	Solicitudes	Aceptaciones	%	Solicitudes	Aceptaciones	%
Informática	26	7	27 %	228	58	25 %
Economía	240	63	26 %	512	112	22 %
Ingeniería	164	52	32 %	972	252	26 %
Medicina	416	99	24 %	578	140	24 %
Veterinaria	338	53	16 %	180	22	12 %
Total	1.184	274	23 %	2.470	584	24 %

Tabla 4.2. Ilustración de la paradoja de Simpson empleando datos de admisión de la Universidad de Cambridge en 1996. Globalmente, la tasa de aceptación era mayor para los hombres. Pero en cada carrera, la tasa de aceptación era mayor para las mujeres.

Pero el ajuste puede producir algunos resultados paradójicos, tal como

muestra un análisis de las tasas de aceptación por género en la Universidad de Cambridge. En 1996 la tasa de aceptación global para estudiar cinco carreras académicas en Cambridge era ligeramente superior para los hombres (24 % de 2.470 candidatos) que para las mujeres (23 % de 1.184 candidatos). Las carreras pertenecen al ámbito de la ciencia, tecnología, ingeniería y matemáticas, que históricamente han estado dominadas por los hombres. ¿Era este un caso de discriminación de género?

Miremos con atención la tabla 4.2. Aunque, globalmente, la tasa de aceptación era mayor para los hombres, la tasa de aceptación en cada carrera individual era mayor para las mujeres. ¿Cómo se puede dar esta aparente paradoja? La explicación es que era más probable que las mujeres optasen por las carreras más populares y por tanto más competitivas, con las menores tasas de aceptación, como Medicina y Veterinaria, y tendían a no optar por la Ingeniería, que tenía una tasa de aceptación mayor. En este caso, por tanto, podríamos concluir que no hay evidencia de discriminación.

Esto se conoce como la paradoja de Simpson, que se produce cuando la aparente dirección de una asociación se invierte como consecuencia de una variable de confusión, lo que lleva, necesariamente, a repensar las lecciones aparentemente extraídas de los datos. A los estadísticos les encanta encontrar ejemplos en la vida real, cada uno de los cuales refuerza la necesidad de ser cautelosos a la hora de interpretar datos observacionales. No obstante, muestra también todo lo que podemos obtener dividiendo los datos de acuerdo con factores que pueden ayudar a explicar asociaciones observadas.

¿Vivir cerca de un Waitrose incrementa en 36.000 libras el valor de su casa?

La afirmación de que vivir cerca de un Waitrose «incrementa en 36.000 libras el valor de su casa» fue recogida acríticamente por los medios británicos en 2017.[71] Pero no se trataba de un estudio sobre el cambio del precio de la vivienda tras la apertura de una tienda, y Waitrose ciertamente no aleatorizó experimentalmente la situación de sus nuevas tiendas: era simplemente una correlación entre los precios de la vivienda y la cercanía a los supermercados, especialmente los de alto nivel, como Waitrose.

La correlación casi con total seguridad refleja la política de Waitrose de abrir tiendas en vecindarios ricos, y es, por tanto, un buen ejemplo de una cadena de causalidad opuesta a lo que se pretende afirmar. Esto se conoce, previsiblemente, como causalidad inversa. Hay ejemplos más serios de ello en los estudios que analizan la relación entre la salud y el consumo de alcohol, que encuentran generalmente que los no bebedores tienen unas tasas de mortalidad sustancialmente mayores que los bebedores moderados. ¿Cómo puede tener esto sentido, dado lo que sabemos sobre el impacto del alcohol sobre el hígado, por ejemplo? Esta relación ha sido en parte atribuida a la causalidad inversa: aquellos que es más probable que mueran no beben porque ya están enfermos (posiblemente debido a que bebieron en exceso en el pasado). Los análisis más cuidadosos excluyen ahora a antiguos bebedores, e, igualmente, ignoran los eventos de salud adversos en los primeros años del estudio, dado que estos podrían deberse a condiciones preexistentes. Incluso tomando estas precauciones, parece mantenerse un beneficio general sobre la salud derivado de un consumo moderado de alcohol, algo que, no obstante, sigue estando muy cuestionado.

Otro ejercicio divertido es intentar inventarse una narrativa de causalidad inversa para cualquier afirmación estadística basada

únicamente en la correlación. Mi favorita es un estudio que encontró una correlación entre el consumo de refrescos gaseosos por parte de los adolescentes estadounidenses y su tendencia a la violencia: aunque un periódico informó de este estudio bajo la rúbrica «Las bebidas gaseosas vuelven violentos a los adolescentes»,[72] ¿quizá simplemente es igual de plausible pensar que la violencia da sed? O más plausiblemente podríamos pensar en algunos factores comunes que podrían influir en ambas variables, como por ejemplo la pertenencia a un determinado grupo de amigos. Las causas potenciales que no medimos se conocen como factores ocultos, dado que permanecen en un segundo plano, no se incluyen en ningún ajuste, y esperan agazapadas para arruinar las conclusiones ingenuas derivadas de datos observacionales.

A continuación, enumero algunos ejemplos de lo fácil que es creer en un vínculo causal, cuando en realidad otros factores están influyendo en los resultados:

→ Muchos niños son diagnosticados de autismo al poco de ser vacunados. ¿Causa autismo la vacunación? No, se trata de eventos que ocurren aproximadamente a la misma edad, e, inevitablemente, en ocasiones coinciden.

→ Del total de personas que mueren cada año, hay una proporción menor de zurdos que en la población en general. ¿Eso significa que los zurdos viven más? No, esto se produce porque las personas que están muriendo ahora nacieron en una época en la que se obligaba a los niños a ser diestros, por lo que simplemente hay menos zurdos mayores.[73]

→ La edad media a la cual mueren los papas es mayor que la de la población general. ¿Significa eso que ser papa te ayuda a vivir más? No, los papas son seleccionados de un grupo que no ha muerto joven (de lo contrario, no serían candidatos a papa).[74]

La enorme cantidad de formas en las que podemos meter la pata podría llevarnos a pensar que nunca podremos extraer resultados causales de algo que no sea un experimento aleatorizado. Pero, quizás irónicamente, esa idea ha sido contrarrestada por la misma persona responsable del primer ensayo clínico aleatorizado moderno.

¿Podemos deducir causalidad a partir de datos observacionales?

Austin Bradford Hill fue un brillante estadístico aplicado británico que estaba a la vanguardia de dos avances científicos de alcance mundial:

diseñó el ensayo clínico sobre la estreptomicina mencionado anteriormente, que esencialmente estableció el estándar para todas las PCA subsiguientes, y, junto con Richard Doll, dirigió la investigación en la década de 1950 que finalmente confirmaría el vínculo entre el consumo de tabaco y el cáncer de pulmón. En 1965 estableció una serie de criterios necesarios para concluir que un vínculo observado entre una variable de exposición y un resultado es causal, siendo esa exposición cualquier cosa, desde elementos químicos a hábitos como el tabaquismo o la falta de ejercicio.

A partir de entonces, estos criterios han sido muy debatidos. La versión que se muestra a continuación fue desarrollada por Jeremy Howick y sus colegas, que distinguieron entre lo que denominan evidencia directa, mecanicista y paralela.[75]

Evidencia directa:

1. El tamaño del efecto es tan grande que no puede ser explicado por una variable de confusión plausible.

2. Hay una proximidad temporal y/o espacial, en el sentido de que la causa precede al efecto y el efecto se produce tras un intervalo plausible, y/o la causa ocurre en el mismo lugar que el efecto.

3. Dosis-respuesta y reversibilidad. El efecto se incrementa a medida que se incrementa la exposición, y la evidencia es aún más sólida si el efecto se reduce al reducirse la dosis.

Evidencia mecanicista:

4. Hay un mecanismo de acción plausible, que puede ser biológico, químico o mecánico, con evidencia externa de una «cadena causal».

Evidencia paralela:

5. El efecto encaja en lo que ya conocemos.

6. Se encuentra el efecto cuando se replica el estudio.

7. Se encuentra el efecto en estudios similares, aunque no idénticos.

Estas directrices podrían ayudar a determinar causalidad a partir de evidencia anecdótica, incluso en ausencia de un estudio aleatorizado. Por ejemplo, se ha observado la aparición de úlceras bucales tras la fricción con la boca de una aspirina, por ejemplo, para evitar el dolor de muelas. El efecto es patente (sigue la directriz número 1), ocurre

cuando se produce fricción (2), es una respuesta plausible a una solución ácida (4), no lo contradice la ciencia y es similar al conocido efecto de la aspirina en la generación de úlceras estomacales (5), y ha sido observado repetidamente en múltiples pacientes (6). De manera que cinco de las siete directrices se cumplen, no habiéndose comprobado las otras dos, por lo que es razonable concluir que se trata de una genuina reacción adversa al medicamento.

Los criterios de Bradford Hill se aplican a conclusiones científicas generales referidas a poblaciones. Pero podríamos también estar interesados en casos individuales, como por ejemplo litigios de responsabilidad civil en los que los tribunales tienen que decidir si una exposición determinada (por ejemplo, al amianto en el lugar de trabajo) ha causado un determinado resultado a una persona específica (por ejemplo, cáncer de pulmón a John Smith). Nunca podrá establecerse con absoluta certeza que el amianto haya sido la causa del cáncer, dado que no puede ser demostrado que el cáncer no se habría producido sin esa exposición. Pero algunos tribunales han aceptado que, en la «balanza de las probabilidades», se considera que hay un vínculo causal directo si el riesgo relativo asociado con la exposición es mayor de dos. Pero ¿por qué dos?

Presumiblemente, el razonamiento que hay detrás de esta conclusión sería el siguiente:

1. Supongamos que, en condiciones normales, de 1.000 hombres como John Smith, 10 desarrollarían cáncer de pulmón. Si el amianto aumenta el riesgo más del doble, entonces si estos 1.000 hombres hubiesen sido expuestos al amianto, quizás 25 habrían desarrollado cáncer de pulmón.

2. De manera que de aquellos expuestos al amianto que han desarrollado cáncer de pulmón, menos de la mitad habría tenido cáncer de pulmón si no hubiesen estado expuestos.

3. De manera que más de la mitad de los cánceres de pulmón en este grupo habrán sido causados por el amianto.

4. Dado que John Smith pertenece a este grupo de personas, entonces en la balanza de probabilidades, su cáncer de pulmón fue causado por el amianto.

Este tipo de argumentación ha llevado a la creación de una nueva área de estudio denominada epidemiología forense, que intenta usar la evidencia derivada de la población para extraer conclusiones sobre

qué puede haber causado los eventos individuales. Esta disciplina, de hecho, ha tenido que crearse para atender a las necesidades de personas que buscan una compensación, pero es en todo caso un área muy exigente para el razonamiento estadístico sobre la causalidad.

Cómo aproximarse al problema de la causalidad todavía es una cuestión conflictiva dentro del campo de la estadística, ya se refiera a los medicamentos o a las orejas grandes, y sin la aleatorización, es difícil extraer conclusiones fiables. Un enfoque imaginativo consiste en aprovecharse del hecho de que muchos genes están distribuidos esencialmente de manera aleatoria en la población, de tal manera que es como si hubiésemos sido aleatorizados en la versión específica de nosotros mismos cuando nacimos. Esto se conoce como la aleatorización mendeliana, por Gregor Mendel, que desarrolló la idea moderna de la genética.[76]

Se han desarrollado otros métodos estadísticos avanzados para ajustar por potenciales variables de confusión y, así, acercarse a una estimación del efecto real de la exposición, basados en gran medida en el análisis de regresión, una de las ideas cruciales de la ciencia estadística. Y por ello tenemos que dar gracias, de nuevo, a la fértil imaginación de Francis Galton.

Resumen

→ La causalidad, en sentido estadístico, significa que cuando intervenimos, las probabilidades de diferentes resultados cambian sistemáticamente.

→ Es difícil establecer causalidad desde la estadística, pero unos estudios aleatorizados bien diseñados son el mejor marco disponible.

→ El método del doble ciego, la intención de tratar y métodos similares han permitido que los ensayos clínicos a gran escala identifiquen efectos moderados pero importantes.

→ Los datos observacionales podrían ocultar factores subyacentes que influyan en la aparente relación observada entre una exposición y un resultado, que podrían ser variables de confusión observadas o factores no observados.

→ Los métodos estadísticos permiten ajustar con base en otros factores, pero siempre hay que enjuiciar hasta qué punto se puede decir que existe causalidad.

[59] «Why Going to University Increases Risk of Getting a Brain Tumour», Mirror Online, 20 de junio de 2016. El artículo original es A. R. Khanolkar et al., «Socioeconomic Position and the Risk of Brain Tumour: A Swedish National Population-Based Cohort Study», Journal of Epidemology and Community Health 70 (2016), pp. 12221228.

[60] T. Vigen, https://www.tylervigen.com/spurious-correlations.

[61] «MRC/BHF Heart Protection Study of Cholesterol Lowering with Simvastatin in 20536 High-Risk Individuals: A Randomised Placebo-Controlled Trial», The Lancet 260 (2002), pp. 7-22.

[62] Por sus siglas en inglés, Heart Protection Study. (N. del T.).

[63] Cholesterol Treatment Trialists' (CTT) Collaborators, «The Effects of Lowering LDL Cholesterol with Statin Therapy in People at Low Risk of Vascular Disease: Meta-Analysis of Individual Data from 27 Randomised Trials», The Lancet 380 (2012), pp. 581-590.

[64] Para personas con mi mismo nivel de riesgo y sin enfermedades previas, estiman una reducción del 25 % del riesgo de un evento vascular serio para una bajada de 1 mmol/l en LDL. Mi LDL disminuyó en 2 mmol/l después de comenzar a tomar estatinas, lo que significa que mi pastilla diaria cambia mi riesgo anual de infarto o ictus por un factor de alrededor de 0,75 × 0,75 = 0,56, o el equivalente a una reducción del riesgo que asumo en un 44 %. Dado que tenía aproximadamente una probabilidad del 13 % de tener un infarto o un ictus en diez años, tomar una estatina reduciría este riesgo al 7 %. Esto significa que la prescripción de mis estatinas ha valido la pena —y más aún si me las tomo—.

[65] Podría parecer aún más sorprendente, y esperanzador, que tanta gente haya decidido participar en uno de estos estudios solamente por el beneficio de futuros pacientes.

[66] Los estudios del Equipo de Análisis del Comportamiento están descritos en: http://www.behaviouralinsights.co.uk/education-and-skills/helping-everyone-reach-their-potential-new-education-results/ y en http://www.behaviouralinsights.co.uk/trial-results/measuring-the-impact-of-body-worn-video-camaras-on-police-behaviorand-criminal-justice-outcomes/.

[67] Por sus siglas en inglés, Study of the Therapeutic Effects of Intercessory Prayer. (N. del T.).

[68] H. Benson *et al.*, «Study on the Therapeutic Effects of Intercessory

Prayer (STEP) in Cardiac Bypass Patients: A Multicenter Randomized Trial of Uncertainty and Certainty of Receiving Intercessory Prayer», American Heart Journal 151 (2006), pp. 934-942.

[69] J. Heathcote, «Why Do Old Men Have Big Ears?», British Medical Journal 311 (1995), https://www.bmj.com/content/311/7021/1668. Véase también «Big Ears: They Really Do Grow as We Age», The Guardian, 17 de julio de 2013.

[70] Desgraciadamente, es poco probable que estas propuestas consigan financiación.

[71] «Waitrose Adds £36000 to House Price», Daily Mail, 29 de mayo de 2017.

[72] «Fizzy Drinks Make Teenagers Violent», Daily Telegraph, 11 de octubre de 2011.

[73] S. Coren y D. F. Halpern, «Left-Handedness: A Marker for Decreased Survival Fitness», Psychological Bulletin 109 (1991), pp. 90-106. Para una crítica a este artículo, véase «Left-Handedness and Life Expectancy», New England Journal of Medicine 325 (1991), pp. 1041-1043.

[74] J. A. Hanley, M. P. Carrieri y D. Serraino, «Statistical Fallibility and the Longevity of Popes: William Farr Meets Wilhem Lexis», International Journal of Epidemiology 35 (2006), pp. 802-805.

[75] J. Hopwick, P. Glasziou y J. K. Aronson, «The Evolution of Evidence Hierarchies: What Can Bradford Hill's «Guidelines for Causation» Contribute?», Journal of the Royal Society of Medicine 102 (2009), pp. 189-194.

[76] La aleatorización mendeliana ha sido empleada, por ejemplo, para examinar la polémica cuestión de si un consumo moderado de alcohol es beneficioso para la salud. Las personas que nunca han consumido alcohol tienden a tener tasas de mortalidad más altas que las que beben un poco, pero no hay acuerdo sobre si eso se debe al alcohol o a que los abstemios son menos saludables por otras razones. Una versión de un gen está asociada con una menor tolerancia al alcohol, y aquellos que lo heredan tienden a beber menos. Los que tienen esa versión del gen y los que no deben ser iguales en todos los otros factores, para que cualquier diferencia sistemática de salud pueda ser atribuida al gen, como si fuera un estudio aleatorizado. Los investigadores han mostrado que las personas que tienen el gen que disminuye la tolerancia al alcohol tienden a ser más saludables, y han

concluido por ello que el alcohol no es bueno para usted. Pero se necesita tener en cuenta supuestos extra para deducir eso, y el debate no ha alcanzado una conclusión definitiva. Véase Y. Cho *et al.*, «Alcohol Intake and Cardiovascular Risk Factors: A Mendelian Randomisation Study», Scientific Reports, 21 de diciembre de 2015.

05

Modelizar relaciones mediante la regresión

Las ideas presentadas en los anteriores capítulos nos permiten visualizar y resumir un único conjunto de números, y también descubrir asociaciones entre parejas de variables. Estas técnicas básicas pueden llevarnos muy lejos, pero los datos con que contamos actualmente son generalmente mucho más complejos. Normalmente habrá una lista de posibles variables relacionadas, una de las cuales será nuestra variable de interés para explicar o predecir, ya sea el riesgo de un individuo de desarrollar cáncer o la población futura de un país. En este capítulo nos ocuparemos de la crucial idea del modelo estadístico, que es una representación formal de las relaciones entre variables, la cual podemos usar para la explicación o la predicción. Esto, inevitablemente, supone introducir algunas ideas matemáticas, pero los conceptos básicos deberían quedar claros aun sin usar álgebra.

Aunque primero volvamos a Francis Galton. Este tenía el típico interés de todo caballero científico victoriano en coleccionar datos, como por ejemplo la sabiduría de las multitudes sobre el peso de un buey. Empleó sus observaciones para hacer previsiones meteorológicas, evaluar la eficacia de las plegarias e incluso comparar la belleza relativa de las mujeres jóvenes en distintas partes del país.[77] También compartía la obsesión de su primo Charles Darwin por la herencia, y se puso a investigar cómo cambian las características personales entre generaciones. Estaba especialmente interesado en la siguiente pregunta:

Usando la estatura de los padres,

¿cómo podemos predecir la estatura de un hijo adulto?

En 1886 Galton reunió unos datos de la estatura de un gran número de padres y sus hijos adultos; los estadísticos de resumen para la mayor parte de los datos se muestran en la tabla 5.1.[78] La muestra de Galton incluía estaturas similares a las de adultos contemporáneos (las estaturas medias de mujeres y hombres adultos en el Reino Unido en 2010 eran de 63 y 69 pulgadas, respectivamente), lo que sugiere que sus sujetos estaban bien alimentados y tenían un alto estatus socioeconómico.

La figura 5.1 muestra un diagrama de dispersión de 465 estaturas de

hijos cruzadas con las de sus padres. Las estaturas de los padres y los hijos están claramente correlacionadas; muestran una correlación de 0,39. ¿Qué tenemos que hacer si queremos predecir la estatura de un hijo a partir de la de un padre? Podríamos empezar escogiendo una línea recta para hacer nuestras predicciones, puesto que eso nos permitiría, para cualquier estatura de un padre, calcular una predicción para la estatura del hijo. Nuestra intuición inmediata podría ser usar la línea diagonal de «igualdad», que supondría que predeciríamos que un hijo adulto tendría la misma estatura que su padre. Pero podemos mejorar esa predicción.

Para cada línea recta que escojamos, cada dato dará lugar a un residuo (las líneas verticales discontinuas en el diagrama), que es el tamaño del error si usásemos la línea para predecir la estatura del hijo a partir de la de su padre. Queremos una línea que haga que estos residuos sean pequeños, y la técnica estándar para ello es escoger la línea de ajuste de mínimos cuadrados, para la cual la suma de los cuadrados de los residuos sea menor.[79] La fórmula para esta línea es muy sencilla (véase el glosario), y fue desarrollada por los matemáticos franceses Adrien-Marie Legendre y Carl Friedrich Gauss, como la predicción que genera el «mejor ajuste» que podemos hacer sobre la estatura de un hijo, sabiendo la de su padre.

	Número	Media	Mediana	Desviación típica
Madres	197	64,0	64,0	2,4
Padres	197	69,3	69,5	2,6
Hijas	433	64,1	64,0	2,4
Hijos	465	69,2	69,2	2,6

Tabla 5.1. Estadísticos de resumen de las estaturas (en pulgadas) de 197 conjuntos de padres y sus hijos adultos recogidas por Galton en 1886. Sesenta y cuatro pulgadas equivalen a 1,63 metros, 69 pulgadas son 1,75 metros. Sin necesidad de hacer un diagrama de dispersión con los datos, la similitud de la media y la mediana sugiere una distribución simétrica.

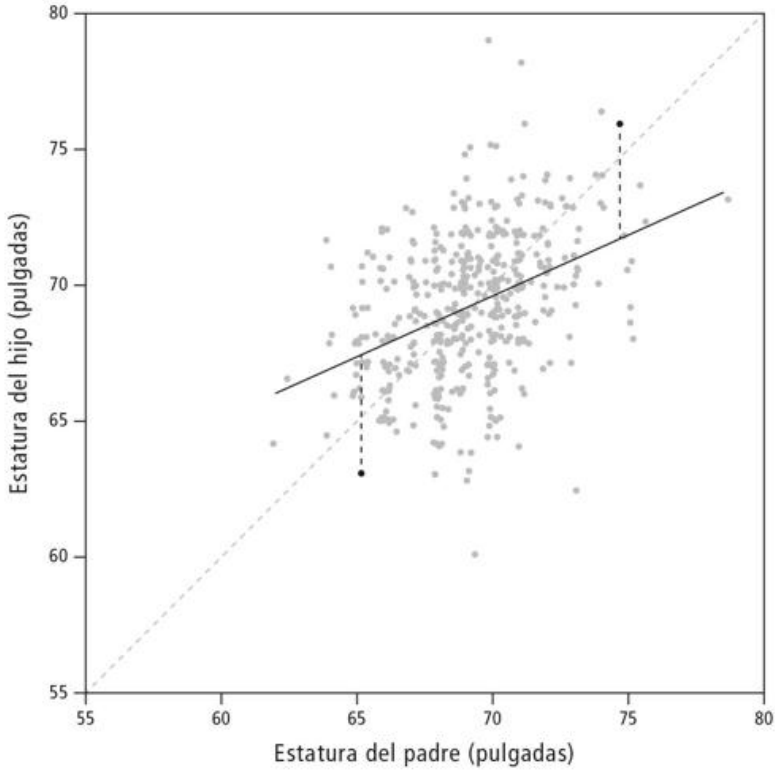


Figura 5.1. Diagrama de dispersión de 465 padres e hijos con los datos de Galton (muchos padres están repetidos, dado que tienen varios hijos). Se ha añadido una fluctuación para separar los puntos, y la línea diagonal discontinua representa la igualdad exacta entre las estaturas de hijos y padres. La línea sólida es la línea estándar del «mejor ajuste». Cada punto da lugar a un «residuo» (la línea discontinua), que es el tamaño del error si usásemos la línea para predecir la estatura del hijo a partir de la del padre.

La línea de predicción de mínimos cuadrados de la figura 5.1 atraviesa la nube de puntos por la mitad, lo que representa los valores medios de las estaturas de padres e hijos, pero no sigue la línea diagonal de «igualdad». Es claramente menor que la línea de igualdad para los padres que son más altos que la media, y mayor que la línea de igualdad para los padres que son más bajos que la media. Esto significa que los padres altos tienden a tener hijos que son ligeramente más bajos que ellos, mientras que los padres más bajos tienen hijos ligeramente más altos. Galton denominó este fenómeno «regresión a la mediocridad», algo que ahora se denomina regresión a la media. Este fenómeno también se da para madres e hijas: las madres más altas tienden a tener hijas que son más bajas que ellas, y las madres más

bajas tienden a tener hijas más altas. Esto explica el origen del término en el título de este capítulo: finalmente, cualquier proceso de ajuste de líneas o curvas a los datos se acabó llamando «regresión».

En el análisis básico de regresión, la variable dependiente es la cantidad que queremos predecir o explicar, normalmente en el eje vertical y en un gráfico; en ocasiones se conoce como variable de respuesta. Y la variable independiente es la cantidad que usamos para realizar la predicción o explicación, normalmente en el eje horizontal x en un gráfico, y que a menudo se conoce como variable explicativa. La pendiente también es conocida como coeficiente de regresión.

La tabla 5.2 muestra las correlaciones entre las estaturas de los padres y de los hijos, y las pendientes de las líneas de regresión.[80] Existe una relación simple entre las pendientes, el coeficiente de correlación de Pearson y las desviaciones típicas de las variables.[81] De hecho, si las desviaciones típicas de las variables independientes y dependientes son las mismas, entonces la pendiente es simplemente el coeficiente de correlación de Pearson, lo que explica su similitud en la tabla 5.2.

	Correlación de Pearson	Pendiente de la regresión de los hijos con respecto a los padres
Madres e hijas	0,31	0,33
Padres e hijos	0,39	0,45

Tabla 5.2. Correlación entre las estaturas de hijos adultos y padres del mismo género, y pendientes de la regresión de la estatura de los hijos con la de los padres.

El significado de estas pendientes depende completamente de nuestras asunciones sobre la relación entre las variables estudiadas. Para datos correlacionados, la pendiente indica cuánto se espera que cambie la variable dependiente, de media, si observamos una diferencia de una unidad en la variable independiente. Por ejemplo, si Alice es una pulgada más alta que Betty, predeciremos que la hija adulta de Alice sea 0,33 pulgadas más alta que la hija adulta de Betty. Por supuesto, no esperaríamos que esta predicción iguale con precisión las verdaderas diferencias de estatura, pero es la mejor estimación que podemos hacer con los datos disponibles.

Si, no obstante, asumimos una relación causal, entonces la pendiente tiene una interpretación muy diferente: es el cambio que esperaríamos en la variable dependiente si interviniésemos y cambiásemos la variable independiente a un valor una unidad superior. Esto no es ni por asomo el caso para las estaturas, dado que no pueden ser alteradas

por medios experimentales, al menos en adultos. Incluso teniendo en cuenta el criterio de Bradford Hill presentado anteriormente, los estadísticos son generalmente reacios a atribuir causalidad a menos que haya habido un experimento, aunque los científicos computacionales Judea Pearl y otros han avanzado mucho en la construcción de modelos de regresión causales a partir de datos observacionales.[82]

Las líneas de regresión son modelos

La línea de regresión que hemos ajustado entre las estaturas de los padres y de los hijos es un ejemplo muy básico de modelo estadístico. La Reserva Federal de Estados Unidos define un modelo como una «representación de algún aspecto del mundo basada en asunciones simplificadas»: esencialmente, algún fenómeno que se representa matemáticamente, generalmente a través de un software informático, para producir una versión «simulada», simplificada, de la realidad.[83]

Los modelos estadísticos tienen dos componentes principales. En primer lugar, una fórmula matemática que expresa un componente determinista, predecible, como, por ejemplo, la línea recta de ajuste que nos permite predecir la estatura del hijo a partir de la del padre. Pero la parte determinista del modelo no va a ser una representación perfecta del mundo observado. Como vimos en la figura 5.1, hay una gran dispersión de las estaturas en torno a la línea de regresión, y la diferencia entre lo que predice el modelo y lo que ocurre realmente es el segundo componente de un modelo y se conoce como error residual —aunque es importante recordar que en los modelos estadísticos, «error» no se refiere a que algo esté mal, sino a la incapacidad inevitable de un modelo de representar lo que observamos—. Por tanto, en resumen, asumimos que:

observación = modelo determinista + error residual

Esta fórmula puede ser interpretada como que, en el mundo de la estadística, lo que vemos y medimos a nuestro alrededor puede ser considerado como la suma de una forma sistemáticamente idealizada matemáticamente, más un componente aleatorio que no puede aún ser explicado. Esta es la idea clásica de señal y ruido.

¿Reducen las cámaras de tráfico los accidentes?

Esta sección contiene una lección simple: el mero hecho de actuar y que tal actuación cambie algo no significa que seamos responsables

del resultado. Los humanos parecen encontrar esta verdad tan sencilla difícil de entender: siempre estamos deseando dar una narración explicativa, y más aún si somos los protagonistas. Por supuesto, en ocasiones esta interpretación es cierta: si accionas un interruptor y la luz se enciende, normalmente eres el responsable. Pero en ocasiones tus acciones son claramente no responsables de un resultado: si no llevas paraguas y llueve, no es culpa tuya (aunque podría parecer que sí). Pero las consecuencias de muchas de nuestras acciones son menos claras. Supongamos que tiene usted un dolor de cabeza, toma una aspirina y su dolor de cabeza desaparece. ¿Cómo sabe si no habría desaparecido de todas formas, aunque no se hubiese tomado la aspirina?

Tenemos una fuerte tendencia psicológica a atribuir el cambio a intervenciones, y esto hace que toda comparación del tipo «antes y después» sea un tanto peligrosa. Un ejemplo clásico son las cámaras de tráfico, que suelen colocarse en lugares donde ha habido accidentes recientemente. Cuando, a continuación, la tasa de accidentes disminuye, esto se atribuye a la presencia de las cámaras. Pero la tasa de accidentes ¿habría disminuido de todas maneras?

Las rachas de buena (o mala) suerte no duran eternamente, y, al final, las aguas vuelven a su cauce (esto también podría considerarse un ejemplo de regresión a la media, igual que el que los padres altos tiendan a tener hijos más bajos). Pero si nos creemos que estas rachas de buena o mala suerte representan un estado de cosas permanente, entonces atribuiremos la reversión a lo normal como la consecuencia de cualquier intervención que hayamos realizado. Quizá esto parezca muy obvio, pero esta idea tan sencilla tiene importantes ramificaciones, como, por ejemplo:

→ Los entrenadores de fútbol que son despedidos tras una racha de derrotas y ven cómo se felicita a sus sucesores por la vuelta a la normalidad.

→ Gestores de fondos cuyo desempeño empeora después de haber recibido un aumento (y quizás obtenido grandes bonificaciones) por un par de años buenos.

→ La «maldición de Sports Illustrated», según la cual los deportistas que aparecen en la portada de una importante revista ven cómo, a continuación, su desempeño cae.

La suerte juega un papel considerable en la posición de los equipos en la clasificación de sus ligas respectivas, y una consecuencia de la

regresión a la media es que esperaríamos que equipos a los que les ha ido bien un año les vaya peor el año siguiente, y aquellos a los que les ha ido mal mejoren, particularmente si los equipos están muy igualados. En cambio, si vemos esta pauta de cambios, podríamos sospechar que lo que está pasando es una regresión a la media en lugar de prestar más atención a la influencia de, digamos, nuevos métodos de entrenamiento.

No son solo los equipos deportivos los que se posicionan en tablas clasificatorias. Tomemos el ejemplo de las tablas del Informe PISA sobre Educación Global, que comparan los sistemas educativos de distintos países en el campo de las matemáticas. Todo cambio de posición en la tabla entre 2003 y 2012 estaba negativamente correlacionado con la posición inicial, lo que quiere decir que los países en las posiciones superiores tienden a empeorar, y los que están abajo tienden a mejorar. La correlación era $-0{,}60$, y hay alguna teoría según la cual si los rankings fuesen completamente aleatorios y todo lo que estuviese operando fuese una regresión a la media, esperaríamos que la correlación fuese de $-0{,}71$, no muy distinta a lo que observamos.[84] Esto sugiere que las diferencias entre países son mucho menores de lo que se dice, y que los cambios en la tabla clasificatoria tienen poco que ver con cambios en la metodología de la enseñanza.

La regresión a la media también opera en los ensayos clínicos. En el último capítulo vimos que se requerían estudios aleatorizados para evaluar adecuadamente los medicamentos, dado que hasta aquellos que están en el grupo de control mostraban mejoras (el llamado efecto placebo). Esto se interpreta normalmente como que tomar una píldora de azúcar (preferiblemente roja) tiene efectos beneficiosos sobre la salud de la gente. Pero gran parte de la mejora percibida en aquellos que no reciben ningún tratamiento activo puede ser una regresión a la media, dado que los pacientes son reclutados para los ensayos cuando muestran síntomas, y muchos de estos síntomas habrían desaparecido en todo caso.

De manera que si queremos conocer el efecto genuino de instalar cámaras de tráfico en puntos negros, entonces deberíamos seguir el enfoque empleado para evaluar medicamentos y dar el paso audaz de instalarlas aleatoriamente. Cuando se han llevado a cabo ese tipo de estudios, la estimación es que dos terceras partes de los supuestos beneficios de las cámaras se deben a la regresión a la media.[85]

Cuando hay más de una variable explicativa

Desde el trabajo inicial de Galton, la idea básica de la regresión se ha expandido considerablemente, ayudada enormemente por la moderna computación. Estos desarrollos incluyen:

→ Tener muchas variables explicativas.

→ Variables explicativas categóricas, no numéricas.

→ Relaciones que no son lineales y que se adaptan flexiblemente a la pauta de los datos.

→ Variables de respuesta que no son continuas, como por ejemplo proporciones o variables contables.

Un ejemplo en el que hay más de una variable explicativa sería ver cómo se relaciona la estatura de un hijo o una hija con la estatura de su padre y de su madre. La dispersión de los datos es ahora tridimensional y es mucho más difícil de representar gráficamente, pero podemos seguir usando la idea de los mínimos cuadrados para derivar la fórmula que predice mejor la estatura de los hijos. Esto se conoce como regresión lineal múltiple.[86]

Cuando solo tenemos una variable explicativa la relación con la variable de respuesta está resumida en la pendiente, que también puede ser interpretada como un coeficiente en una ecuación de regresión; esta idea puede generalizarse a más de una variable explicativa.

Los resultados para el caso de las familias de Galton se muestran en la tabla 5.3. ¿Cómo podemos interpretar los coeficientes de la tabla? En primer lugar, son parte de una fórmula que podría emplearse para predecir la estatura de los hijos adultos para una madre y padre específicos.[87] Pero también ilustran la idea de ajuste de una relación aparente, incluyendo una tercera variable de confusión.

Por ejemplo, vimos en la tabla 5.2 que la pendiente de la regresión de la estatura de las hijas sobre la estatura de las madres era de 0,33 —recordemos que la pendiente de una línea ajustada para un gráfico de dispersión es solo otro nombre aplicado al coeficiente de la regresión —. La tabla 5.3 muestra que si controlamos por la estatura del padre, este coeficiente se reduce a 0,30. Cuando predecimos la estatura de un hijo, el coeficiente de regresión para el padre se reduce igualmente de 0,45 en la tabla 5.2 a 0,41 en la tabla 5.3, cuando se controla por la estatura de la madre. De manera que la estatura de uno de los progenitores tiene una asociación ligeramente negativa con la estatura

de su hijo adulto cuando se controla por el efecto del otro progenitor. Esto se debe al hecho de que las mujeres altas tienden a casarse con hombres altos, por lo que la estatura de cada progenitor no es un factor completamente independiente. En líneas generales, los datos sugieren que una pulgada de diferencia en la estatura de un padre está asociada con una mayor diferencia en la estatura de un hijo adulto que una diferencia de una pulgada en la estatura de la madre. La regresión múltiple se usa en muchas ocasiones cuando los investigadores están interesados en una determinada variable explicativa, pero se necesita «controlar por» otras variables para que el modelo no esté desequilibrado.

Variable dependiente	Constante (estatura media de los hijos)	Coeficiente de la regresión múltiple para la estatura de la madre	Coeficiente de la regresión múltiple para la estatura del padre
Estatura de la hija	64,1	0,30	0,40
Estatura del hijo	69,2	0,33	0,41

Tabla 5.3. Resultados de una regresión lineal múltiple que relaciona la estatura de los hijos adultos con la de la madre y el padre. La «constante» es la estatura media de los hijos (tabla 5.1). Los coeficientes de la regresión múltiple indican el cambio predicho en la estatura de los hijos adultos para cada cambio de una pulgada en la estatura media de los padres.

Volvamos al estudio sueco sobre los tumores cerebrales que vimos en el capítulo 4 como ejemplo de una interpretación inadecuada de la causalidad por parte de los medios. Un análisis de regresión tendría la tasa de tumores como variable dependiente o de respuesta, y la educación como variable independiente, o explicativa, de interés. Otros factores incluidos en la regresión serían la edad en el momento del diagnóstico, el año, la región de Suecia, el estado civil y los ingresos, cada uno de los cuales se consideraría una potencial variable de confusión. La inclusión de estas variables de confusión se hace para establecer cuál es la auténtica relación entre educación y tumores cerebrales, pero es una estrategia que nunca es del todo satisfactoria. Siempre nos quedará la sospecha de que pueda haber algún factor no observado que se nos escape, como por ejemplo que las personas con educación superior sean las que procuren tener una mejor salud, y por ello un número mayor de diagnósticos.

En un estudio aleatorizado, no habría necesidad de controlar por variables de confusión, dado que la asignación aleatoria garantiza que todos los factores, excepto el tratamiento principal, estarán

equilibrados entre los grupos. Pero los investigadores a menudo llevan a cabo de todas maneras un análisis de regresión, por si acaso se ha colado algún tipo de desequilibrio.

Distintos tipos de variables de respuesta

No todos los datos son medidas continuas como la estatura. En gran parte de los análisis estadísticos, las variables dependientes pueden ser, por ejemplo, el porcentaje de eventos que ocurren o no (por ejemplo, la proporción de personas que sobreviven a la cirugía), un recuento del número de eventos (por ejemplo, cuántos cánceres se producen al año en un área determinada) o el tiempo transcurrido antes de que se produzca un evento (por ejemplo, los años de supervivencia después de la cirugía). Cada tipo de variable dependiente tiene su forma propia de regresión múltiple, con su correspondiente interpretación de los coeficientes estimados.[88]

Consideremos los datos de cirugía cardiaca infantil que discutimos en el capítulo 2. La figura 2.5(a) mostraba el porcentaje de supervivencia a la cirugía y el número de casos tratados en cada hospital entre 1991 y 1995. El diagrama de dispersión se muestra de nuevo en la figura 5.2, con una línea de regresión ajustada sin tener en cuenta el dato atípico correspondiente a Bristol.

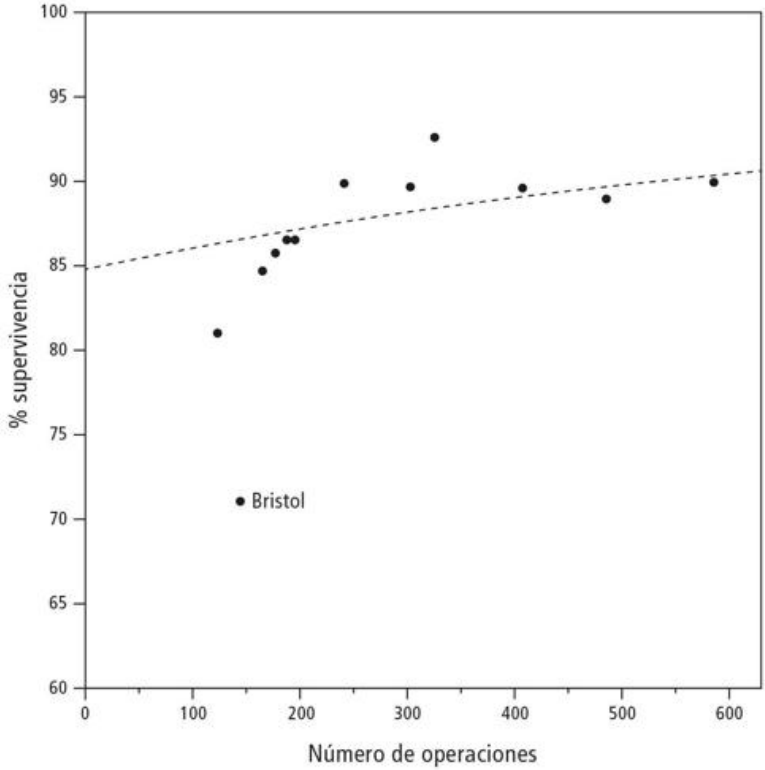


Figura 5.2. Modelo de regresión logística ajustada a los datos de cirugía cardiaca infantil para menores de 1 año en hospitales del Reino Unido entre 1991 y 1995. Los hospitales que tratan a más pacientes tienen mejores tasas de supervivencia. La línea es parte de una curva que nunca alcanza el 100 %, y está ajustada ignorando los datos atípicos correspondientes a Bristol.

Aunque podríamos haber ajustado una regresión lineal a través de estos puntos, una extrapolación inocente nos indicaría que si un hospital ha tratado un enorme número de casos, se predeciría que su supervivencia sería mayor del 100 %, lo cual es absurdo. Por ello, se ha desarrollado una regresión aplicada a porcentajes, denominada regresión logística, cuya curva no puede superar el 100 % o quedar por debajo del 0 %.

Incluso sin tener en cuenta los datos de Bristol, los hospitales con más pacientes tienen mejores tasas de supervivencia, y el coeficiente de regresión logística (0,001) significa que se espera que la tasa de mortalidad sea alrededor de un 10 % (relativamente) menor por cada 100 operaciones realizadas por un hospital a menores de 1 año durante un período de 4 años.[89] Por supuesto, por usar lo que ahora

es más bien un cliché, correlación no significa causalidad, y no podemos concluir que una mayor productividad sea la causa de un mejor desempeño: como mencionamos antes, podría tratarse incluso de causalidad inversa, en el sentido de que los hospitales con una buena reputación atraen a más pacientes.

Este descubrimiento generó mucha controversia cuando fue anunciado en 2001, y ha contribuido a las prolongadas y aún irresueltas discusiones sobre cuántos hospitales en el Reino Unido deberían llevar a cabo esta forma de cirugía.

Más allá de los modelos básicos de regresión

Las técnicas referidas en este capítulo han funcionado sorprendentemente bien desde su introducción hace más de un siglo. Pero tanto la disponibilidad de una gran cantidad de datos como el incremento extraordinario de la capacidad computacional han permitido el desarrollo de modelos mucho más sofisticados. A grandes rasgos, distintos investigadores han adoptado cuatro estrategias principales de modelización:

→ Representaciones matemáticas simples para asociaciones, como los análisis de regresión lineal de este capítulo, que tienden a ser las favoritas de los estadísticos.

→ Modelos deterministas complejos basados en la comprensión científica de un proceso físico, como los empleados en la previsión del tiempo, cuya intención es representar realistamente mecanismos subyacentes, y que son desarrollados generalmente por matemáticos estadísticos.

→ Algoritmos complejos usados para tomar decisiones o predicciones que se han derivado de un análisis de un número enorme de casos pasados, por ejemplo, para recomendar libros que te gustaría comprar en una tienda online, y que vienen del mundo de la informática y el aprendizaje automático. Son a menudo «cajas negras» en el sentido de que puede que proporcionen buenas predicciones, pero su estructura interna es un tanto inescrutable (algo que veremos en el siguiente capítulo).

→ Modelos de regresión que afirman alcanzar conclusiones causales, los favoritos de los economistas.

Son estas unas generalizaciones enormes, y afortunadamente las barreras profesionales están siendo derribadas, y, como veremos más

adelante, se está desarrollando un modelo más ecuménico. Pero sea cual sea la estrategia adoptada, siempre surgen las mismas cuestiones cuando se estima un modelo.

Una buena analogía sería decir que un modelo es como un mapa, más que como el propio territorio. Y ya sabemos que algunos mapas son mejores que otros: uno simple podría ser suficiente para conducir entre ciudades, pero necesitamos algo más detallado cuando paseamos por el campo. El estadístico británico George Box se hizo famoso por su breve pero impagable aforismo: «Todos los modelos están equivocados, algunos son útiles». Esta afirmación tan sucinta se basaba en toda una vida dedicada a aplicar los conocimientos estadísticos a los procesos industriales, lo que llevó a Box a apreciar tanto el poder de los modelos como el peligro de creer demasiado en ellos.

Pero estas precauciones son fáciles de olvidar. Una vez que un modelo se acepta, y especialmente cuando se escapa al control de aquellos que lo han creado y entienden sus limitaciones, empieza a actuar como si fuera un oráculo. La exagerada confianza en unos modelos financieros complejos empleados para determinar el riesgo de, por ejemplo, paquetes de hipotecas, ha sido considerada en gran medida como responsable de la crisis financiera de 2007-2008. Estos modelos asumían solo una correlación moderada entre los impagos de hipotecas, y funcionaban bien mientras el mercado de la vivienda estaba en plena expansión. Pero cuando las condiciones cambiaron y comenzaron los impagos de las hipotecas, tendieron a hacerlo de forma masiva: el modelo subestimó enormemente los riesgos, debido a que las correlaciones resultaron ser mucho mayores de lo que se suponía. Los altos directivos sencillamente no eran conscientes de las frágiles bases sobre las que se construían estos modelos, se habían olvidado de que los modelos son simplificaciones del mundo real —son los mapas, no el territorio—. El resultado fue una de las peores crisis económicas globales de la historia.

Resumen

→ Los modelos de regresión proporcionan una representación matemática de la relación entre un conjunto de variables explicativas y una variable de respuesta.

→ Los coeficientes de un modelo de regresión indican cuánto se espera que cambie la respuesta cuando se observa un cambio en la variable explicativa.

→ La regresión a la media se produce cuando las respuestas más extremas revierten a largo plazo a niveles más cercanos a la media, dado que la generación previa de esos valores extremos era puramente aleatoria.

→ Los modelos de regresión pueden incorporar distintos tipos de variables de respuesta, variables explicativas y relaciones no lineales.

→ Hay que ser cuidadoso a la hora de interpretar los modelos, no debemos hacerlo demasiado literalmente: «Todos los modelos están equivocados, pero algunos son útiles».

[77] Según Galton, «Londres está primero en bellezas; Aberdeen, último».

[78] M. Friendly et al., «HistData: Data Sets from the History of Statistics and Data Visualization», 2018, https://CRAN.R-project.org/package=HistData.

[79] Sería posible ajustar una línea que minimice la suma de los valores absolutos de los residuos, en lugar de la suma de sus cuadrados, pero esto sería casi imposible sin un ordenador moderno.

[80] Por ejemplo, predeciríamos la estatura de una hija mediante la fórmula: estatura media de todas las hijas + 0,33 × (estatura de la madre – estatura media de todas las madres).

[81] Véase la entrada de «mínimos cuadrados» en el glosario.

[82] J. Pearl y D. Mackenzie, The Book of Why: The New Science of Cause and Effect, Basic Books, 2018, p. 471.

[83] Para una discusión fascinante de los riesgos de la modelización, véase A. Aggarwal et al., «Model Risk—Daring to Open Up the Black Box», British Actuarial Journal 21:2 (2016), pp. 229-296.

[84] Lo que estamos diciendo, básicamente, es que los cambios estarán correlacionados con una medida de partida, incluso aunque no haya un cambio real en el proceso subyacente. Podemos expresar esto matemáticamente. Supongamos que tomo una observación de manera aleatoria de una distribución poblacional, llamémosla X. Luego tomo otra observación independiente de la misma distribución, llamémosla Y, y miro la diferencia entre ambas: Y – X. Un hecho de lo más sorprendente es que la correlación entre su diferencia, Y − X, y la primera medida, X, es $-(1/\sqrt{2}) = -0{,}71$, con independencia de la

forma de la distribución poblacional subyacente. Por ejemplo, si una mujer ha tenido un hijo, y después su amiga tiene también uno, y deciden ver hasta qué punto el bebé de la amiga pesa más, calculando el peso del segundo menos el peso del primero, entonces esta diferencia tiene una correlación de −0,71 con el peso del primer bebé. Esto es debido a que si el primer niño no pesa mucho, esperaríamos que el segundo, aleatoriamente, pese más, de manera que la diferencia sería positiva. Y si el primer niño pesa mucho, entonces esperaríamos que la diferencia entre los pesos sea negativa.

[85] L. Mountain, «Safety Cameras: Stealth Tax or Life-Savers?», Significance 3 (2006), pp. 111-113.

[86] Lo de «lineal» se refiere a que esta ecuación consiste en una suma ponderada de las variables explicativas. La ponderación es por sus coeficientes de regresión, y esto es lo que se conoce como un modelo lineal.

[87] Las variables explicativas han sido estandarizadas restándoles sus valores muestrales medios. De manera que para predecir la estatura de un hijo, usaríamos la fórmula: 69,2 + 0,33 (estatura de la madre – estatura media de las madres) + 0,41 (estatura del padre – estatura media de los padres).

[88] La tabla que se muestra a continuación muestra las formas de regresión múltiple usadas para distintos tipos de variable dependiente. Cada una de ellas da como resultado un coeficiente de regresión estimado para cada variable explicativa.

Tipo de variable dependiente	Tipo de regresión	Interpretación de los coeficientes
Variables continuas	Lineal múltiple	Pendiente
Eventos o porcentajes	Logística	Log (razón de probabilidad)
Recuento	Poisson	Log (razón de predicción)
Duración de supervivencia	Cox	Log (riesgo relativo)

[89] El coeficiente de la regresión logística significa que se estima que el logaritmo de la probabilidad de mortalidad disminuya en un 0,001 por paciente extra tratado por año, de manera que disminuye en 0,1 por cada 100 pacientes extra. Esto corresponde a un riesgo un 10 % menor.

06

Algoritmos, analítica y predicción

Hasta ahora el libro se ha centrado en el análisis de cómo la ciencia estadística puede ayudarnos a entender el mundo, ya sea estableciendo el daño potencial de comer bocadillos de beicon o la relación entre la estatura de los padres y los hijos. Esto es, esencialmente, investigación científica para desentrañar qué es lo que está pasando realmente y qué, en términos del último capítulo, es solo un error residual que debe tratarse como una variabilidad inevitable que no puede ser modelizada.

Pero las ideas básicas de la ciencia estadística aún se mantienen cuando intentamos solventar un problema práctico más que científico. El deseo de encontrar una pauta en el ruido es igual de relevante cuando lo que queremos es un método que nos ayude en las decisiones a las que nos enfrentamos en nuestras vidas diarias. El objetivo de este capítulo es mostrar cómo pueden afrontarse esos problemas prácticos mediante datos previos para generar un algoritmo, una fórmula mecánica que produzca automáticamente una respuesta para cada caso, con ninguna (o una mínima) intervención humana adicional; esto es, en efecto, «tecnología» más que ciencia.

Un algoritmo tal tiene dos objetivos amplios:

→ Clasificación (también conocida como discriminación o aprendizaje supervisado): para decirnos a qué tipo de situación nos enfrentamos. Por ejemplo, las preferencias de un cliente en Internet o si ese objeto en el campo de visión de un robot es un niño o un perro.

→ Predicción: para decirnos qué va a pasar. Por ejemplo, qué tiempo hará la semana que viene, cómo evolucionará mañana el precio de unas acciones, qué productos podría comprar un cliente o si ese niño se va a cruzar en el camino de nuestro coche automático.

Aunque estas tareas difieren en si se refieren al presente o al futuro, ambas tienen la misma naturaleza subyacente: utilizan un conjunto de observaciones relevantes para una determinada situación actual, y extraen de ellas una conclusión relevante. Este proceso se ha denominado analítica predictiva, pero bordea el territorio de la inteligencia artificial (IA), en la cual los algoritmos encarnados en máquinas se usan o bien para llevar a cabo tareas que normalmente requerirían intervención humana, o para proporcionar consejo experto

a humanos.

La IA «estrecha» se refiere a los sistemas que pueden llevar a cabo tareas cuidadosamente prescritas. Existen algunos ejemplos extraordinariamente exitosos de este tipo de IA basados en aprendizaje automático, que implica desarrollar algoritmos a través de análisis estadísticos de grandes conjuntos de ejemplos históricos. Algunos éxitos notables son los sistemas de reconocimiento de voz en teléfonos, tabletas y ordenadores; programas como el traductor de Google, que sabe poco de gramática, pero ha aprendido a traducir textos de un archivo publicado inmenso; y el software informático de visión que usa imágenes previas para «aprender» a identificar, por ejemplo, rostros en fotografías u otros coches en el campo visual de vehículos automáticos. También se han producido progresos espectaculares en sistemas de juegos, como el software DeepMind, que aprendió las normas de los juegos de ordenador y se convirtió en un jugador experto, derrotando a campeones mundiales de ajedrez y de go, mientras que Watson, de IBM, ha derrotado a competidores humanos en concursos de cultura general. Estos sistemas no fueron planeados para codificar la experiencia y el conocimiento humanos. Comenzaron con un vasto número de ejemplos y aprendieron mediante prueba y error como un niño inocente, incluso jugando a distintos juegos.

Pero, una vez más, debemos enfatizar que se trata de sistemas tecnológicos que usan datos del pasado para responder a cuestiones prácticas inmediatas, y no sistemas científicos que intentan comprender cómo funciona el mundo: tienen que ser juzgados únicamente por cómo llevan a cabo esa tarea más bien limitada, y, aunque la forma de los algoritmos de aprendizaje podría proporcionar algunas ideas nuevas, no hay que esperar que tengan imaginación o conocimientos sobrehumanos sobre la vida cotidiana. Esto requería IA «general», algo que está más allá de los objetivos de este libro, y, al menos en la actualidad, de la capacidad de las máquinas.

Desde que Edmund Halley desarrollase las fórmulas para el cálculo de seguros y anualidades en la década de 1690, la ciencia estadística se ha ocupado de producir algoritmos para ayudar a las decisiones humanas. El moderno desarrollo de la ciencia de los datos continúa con esa tradición, pero lo que ha cambiado en los últimos años es la escala de los datos y los imaginativos productos que se están desarrollando a partir de ellos: los llamados «macrodatos».

Los datos pueden ser «macro» de dos formas distintas. En primer lugar, por el número de ejemplos en la base de datos, que pueden ser

personas individuales, pero también estrellas en el cielo, escuelas, viajes en coche o entradas en redes sociales. El número de ejemplos es indicado a menudo mediante la letra n, y al comienzo de mi carrera, n era «grande» si tenía más de 100, pero ahora hay datos basados en muchos millones o miles de millones de casos.

La otra forma en la que los datos pueden ser «grandes» es midiendo muchas características en cada ejemplo. Esta cualidad se conoce normalmente como p, referido, quizá, a parámetros. Pensando de nuevo en el comienzo de mi carrera en la estadística, p solía ser menor de 10; quizá entonces conocíamos solo unos pocos rasgos de la historia médica de los individuos. Pero entonces comenzamos a tener acceso a millones de genes de una persona, y la genómica se convirtió en un problema de n pequeño y p grande, en el que contábamos con una cantidad enorme de información sobre un número relativamente pequeño de casos.

Y ahora estamos en la era de los problemas con un gran n y una gran p, en los cuales hay un vasto número de casos, cada uno de los cuales puede ser muy complejo; piénsese en los algoritmos que analizan todas las entradas, los «me gusta» y los «no me gusta» de cada uno de los miles de millones de miembros de Facebook que deciden qué tipo de anuncios y noticias consumir.

Todo ello son desafíos de lo más excitantes que han atraído a oleadas de nuevos analistas a la ciencia de los datos. Pero, por referirnos una vez más a la advertencia que hicimos al comienzo del libro, esta cantidad apabullante de datos no habla por sí sola. Los datos tienen que ser manejados con cautela si queremos evitar los muchos problemas potenciales derivados de un uso cándido de estos. Veremos algunos desastres clásicos en este capítulo, pero primero necesitamos considerar el problema fundamental de cómo cocinar los datos para obtener algo que resulte útil.

Encontrar pautas

Una estrategia para manejar un número excesivo de casos es identificar grupos que sean similares, un proceso conocido como agrupamiento o aprendizaje no supervisado, porque tenemos que aprender acerca de unos grupos que no sabemos por anticipado si existen. Encontrar estos grupos homogéneos puede ser un fin en sí mismo, por ejemplo, identificando grupos de personas con un número similar de «me gusta» y «no me gusta», que después pueden ser caracterizados, asignados a una etiqueta y ser útiles para crear algoritmos para clasificar casos futuros. A partir de ahí se pueden

hacer recomendaciones de películas, anuncios o propaganda política, dependiendo de las motivaciones de la gente que crea el algoritmo, a los grupos que han sido identificados.

Antes de continuar con la creación de un algoritmo para clasificación o predicción, podríamos tener que reducir los datos en bruto de cada caso a una dimensión manejable, porque tal vez tengamos una p excesivamente grande, es decir, demasiadas características por cada caso. Este proceso se conoce como ingeniería de características. Piénsese en el número de medidas referidas a un rostro humano, que podrían tener que ser reducidas a un número limitado de características importantes que puedan usarse para un software de reconocimiento facial que compare una fotografía con otras contenidas en una base de datos. Se podrían identificar medidas carentes de valor predictivo o clasificatorio por medio de visualización de datos o métodos de regresión y luego descartarlas, o se podría reducir el número de características mediante la creación de medidas compuestas que condensen la mayoría de la información.

Desarrollos recientes en modelos extremadamente complejos, como los englobados bajo la etiqueta de aprendizaje profundo, sugieren que este estadio inicial de reducción de los datos podría no ser necesario y que todos los datos en bruto podrían ser procesados en un único algoritmo.

Clasificación y predicción

En la actualidad, podemos disponer de un número increíblemente amplio de métodos alternativos para construir algoritmos de clasificación y predicción. Los investigadores solían promover métodos originarios de sus propios ámbitos profesionales: por ejemplo, los estadísticos preferían los modelos de regresión, mientras que los científicos computacionales preferían la lógica basada en reglas o «redes neuronales», que son formas alternativas de intentar imitar la condición humana. La implementación de cualquiera de estos métodos requería conocimientos y software especializados, pero ahora hay programas que permiten un menú de opciones técnicas y, así, favorecen un enfoque menos parcial donde el desempeño es más importante que la filosofía que hay detrás de la modelización.

Tan pronto como el desempeño práctico de los algoritmos empezó a ser medido y comparado, la gente invariablemente se volvió competitiva, una consecuencia de lo cual es que existen en la actualidad campeonatos de ciencia de datos en plataformas como Kaggle.com. Una organización comercial o académica proporciona un

conjunto de datos para que los descarguen los competidores: los desafíos a los que se tienen que enfrentar son, por ejemplo, la detección de ballenas a partir de grabaciones de sonido, el cálculo de la materia oscura en datos de astronomía o la predicción de admisiones hospitalarias. En cada caso se proporciona a los competidores un conjunto de datos sobre los que construir su algoritmo y un test para decidir su desempeño. Una competición particularmente popular, con miles de equipos participantes, consiste en producir un algoritmo para el siguiente desafío:

¿Podemos predecir qué pasajeros sobrevivieron al hundimiento del Titanic?

En su viaje inaugural, el Titanic chocó contra un iceberg y se hundió lentamente en la noche del 14-15 de abril de 1912. De los más de 2.200 pasajeros y tripulación a bordo, solo 700 pudieron subirse a los botes salvavidas y salvarse. Los subsiguientes estudios y relatos de ficción se han centrado en el hecho de que la probabilidad de subirse a un bote salvavidas dependiese del tipo de billete que se tuviera.

Un algoritmo para predecir la supervivencia podría a primera vista ser un Problema atípico dentro del marco del ciclo PPDAC estándar, dado que es muy poco probable que esta situación se repita, por lo que no tendrá ningún valor en el futuro. Pero un determinado individuo me dio una motivación. En 1912 Francis William Somerton dejó Ilfracombe, en el norte de Devon, cerca de donde yo nací y crecí, para irse a Estados Unidos a hacer fortuna. Dejó en Inglaterra a su mujer y a su joven hija, y compró un billete de tercera clase por ocho libras y un chelín para el recién estrenado Titanic. Nunca llegaría a Nueva York —su lápida está en el cementerio de Ilfracombe (figura 6.1)—. Un algoritmo predictivo preciso sería capaz de decirnos si Francis Somerton tuvo mala suerte o si sus probabilidades de sobrevivir eran más bien escasas.

El Plan es reunir los datos disponibles e intentar una serie de técnicas distintas para producir algoritmos que predigan quién sobrevivió —esto podría considerarse más un problema de clasificación que de predicción, dado que los eventos ya se han producido—. Los datos comprenden la información públicamente disponible de 1.309 pasajeros del Titanic: las potenciales variables predictivas son su nombre completo, título, género, edad, clase en la que viajaban (primera, segunda, tercera), cuánto pagaron por su billete, si eran parte de una familia, dónde embarcaron (Southampton, Cherburgo, Queenstown), así como datos limitados sobre algunos camarotes.[90] La variable de respuesta es un indicador de si sobrevivieron (1) o no

(0).

HENRY SOMERTON
DIED AUGUST 22ND 1913
AGED 74 YEARS.
ALSO OF
FRANCIS WILLIAM
WHO PERISHED
THE TITANIC DISASTER

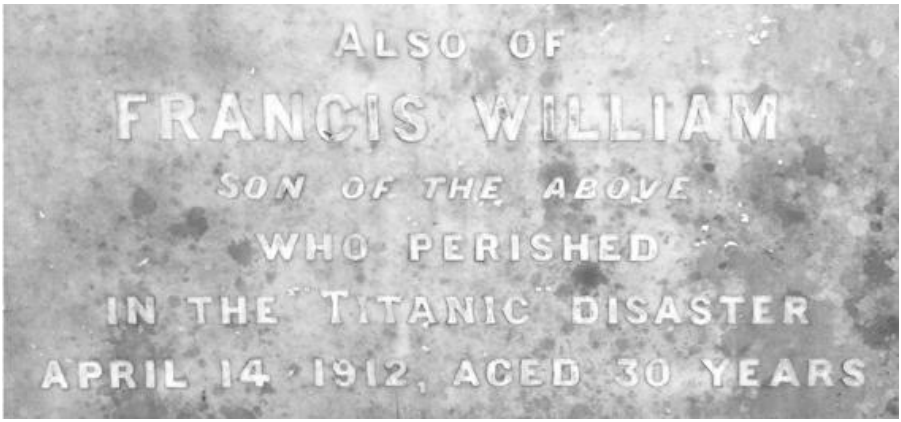


Figura 6.1. Lápida de Francis William Somerton en el cementerio de Ilfracombe. Reza lo siguiente: «También de Francis William, hijo del anterior, que pereció en el desastre del "Titanic" el 14 de abril de 1912, a la edad de 30 años».

Para el Análisis, es crucial dividir los datos en un conjunto de entrenamiento, que se emplea para construir el algoritmo, y un conjunto de validación, que se deja a un lado y se usa exclusivamente para evaluar el desempeño —sería hacer trampas mirar el conjunto de validación antes de que tengamos listo nuestro algoritmo—. Como en la competición de Kaggle, tendremos una muestra aleatoria de 897 casos como nuestro conjunto de entrenamiento, mientras que los otros 442 individuos constituirán nuestro conjunto de validación.

Se trata de una base de datos real, y, por ello, bastante desordenada, por lo que hay que realizar algún tipo de preprocesado. Hay dieciocho pasajeros para los cuales se carece de información sobre el precio de su billete, así que asumimos que han pagado el precio mediano para la clase en la que viajan. Se han sumado hijos y padres para crear una única variable que resuma el tamaño de la familia. Se han tenido que simplificar los títulos: «Mlle» y «Ms» han sido recodificados como «Miss», «Mme» como «Mrs», y una variedad de otros títulos han sido todos ellos codificados como «títulos raros».[91]

Debería quedar claro que, aparte de los conocimientos en codificación requeridos, se necesita buen juicio y conocimiento del contexto, simplemente para que los datos estén listos para ser analizados. Por ejemplo, para usar cualquier información disponible sobre los camarotes para determinar la posición de un individuo en el buque. No cabe duda, en todo caso, de que lo podría haber hecho mejor.

La figura 6.2 muestra la proporción de las distintas categorías de

pasajeros que sobrevivieron, para los 897 pasajeros en el conjunto de entrenamiento. Todas estas características tienen una capacidad predictiva propia, con mayores tasas de supervivencia entre pasajeros que viajan en una clase mejor, las mujeres, los niños, los que pagaron más por su billete, los que tenían una familia de tamaño moderado y los que tenían el título de «Mrs», «Miss» o «Master». Todo esto es coherente con lo que ya sospechábamos.

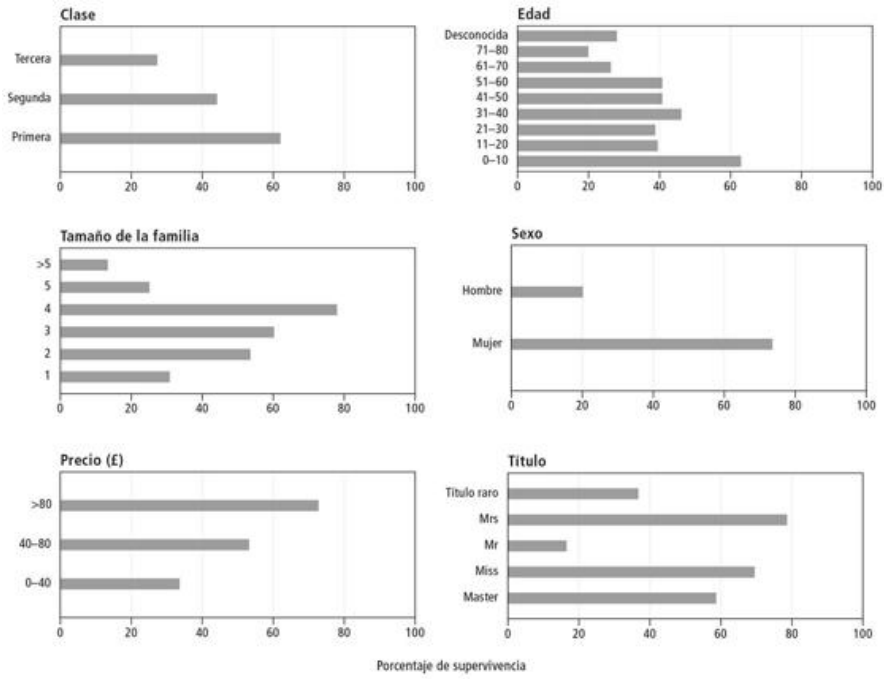


Figura 6.2. Estadísticas de resumen sobre supervivencia para el conjunto de entrenamiento de 897 pasajeros del Titanic, con los porcentajes de las distintas categorías de supervivientes.

Pero estas características no son independientes. Los pasajeros en clases superiores presumiblemente pagaron más por sus billetes, y se esperaría que viajasen con menos hijos que los emigrantes más pobres. Muchos hombres viajaban solos. Y la codificación específica podría ser importante: ¿deberíamos considerar la edad como una variable categórica, dividida en las categorías mostradas en la figura 6.2, o como una variable continua? Los competidores han dedicado mucho tiempo a estudiar estas características y codificarlas para extraer la máxima información, pero nosotros, en lugar de ello, pasaremos directamente a nuestras predicciones.

Supongamos que hacemos la (demostrablemente incorrecta)

predicción de que «nadie sobrevivió». Dado que un 61 % de los pasajeros murieron, podríamos acertar un 61 % de los casos en el conjunto de entrenamiento. Si usásemos la regla de predicción ligeramente más compleja, «todas las mujeres sobreviven y ningún hombre sobrevive», clasificaríamos correctamente al 78 % del conjunto de entrenamiento. Estas reglas sencillas sirven como puntos de partida sobre los que medir cualquier mejora obtenida a partir de algoritmos más sofisticados.

Árboles de clasificación

Un árbol de clasificación es quizás la forma más simpe de algoritmo, puesto que consiste en una serie de preguntas de sí o no, la respuesta a las cuales decide la siguiente pregunta, hasta que se alcanza una conclusión. La figura 6.3 muestra un árbol de clasificación para los datos del Titanic, en el cual los pasajeros están asignados al resultado mayoritario al final de cada rama. Es fácil ver los factores que han sido escogidos, así como la conclusión final. Por ejemplo, Francis Somerton tenía el título de «Mr» en la base de datos, y por ello se sitúa en la primera rama a la izquierda. El extremo de esta rama contiene el 58 % del conjunto de entrenamiento, y de ellos un 16 % sobrevivió. Podríamos, por tanto, afirmar, basándonos en esta limitada información, que Somerton tenía un 16 % de probabilidades de sobrevivir. Nuestro algoritmo simple identifica dos grupos con más de un 50 % de supervivientes: mujeres y niños en primera y segunda clase (en tanto no tuviesen un título raro), de los cuales un 93 % sobrevivió. Y las mujeres y los niños en tercera clase, siempre que viniesen de familias pequeñas, en cuyo caso el 60 % sobrevivió.

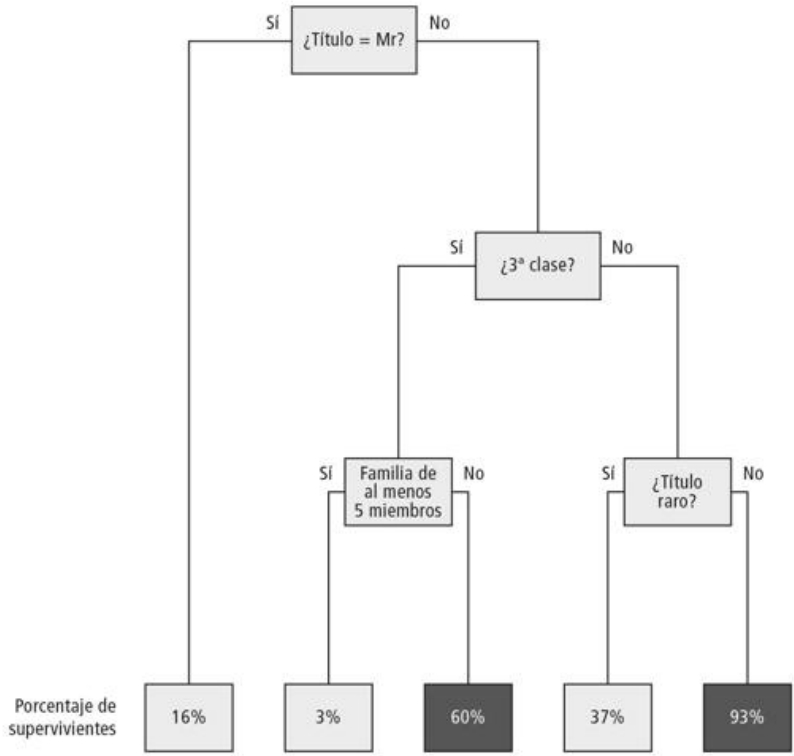


Figura 6.3. Árbol de clasificación de datos del Titanic en el cual una secuencia de preguntas lleva a un pasajero al extremo de una rama, momento en el cual se predice que sobrevivirá si la proporción de personas similares en el conjunto de supervivencia es mayor de un 50 %; estos porcentajes de supervivencia se muestran en la parte inferior del árbol. Las únicas personas que se prevé que sobrevivan son mujeres y niños de familias pequeñas en tercera clase, y todas las mujeres y niños en primera y segunda clase, siempre que no tuviesen títulos raros.

Antes de ver cómo se construye un árbol de ese tipo, necesitamos decidir qué medidas de desempeño usar en nuestra competición.

Evaluación del desempeño de un algoritmo

Para que los algoritmos puedan competir por ver cuál es el más «preciso», alguien tiene que decidir qué se entiende por «preciso». En el desafío Kaggle sobre el Titanic, sería simplemente el porcentaje de pasajeros en el conjunto de entrenamiento que son clasificados correctamente, de manera que una vez que los competidores han creado su algoritmo, suben sus predicciones sobre la variable de respuesta al conjunto de validación, y Kaggle mide su precisión.[92]

Presentaremos los resultados de todo el conjunto de validación a la vez (recalcando que no son los mismos que el conjunto de validación de Kaggle).

El árbol de clasificación mostrado en la figura 6.3 tiene una precisión del 82 % aplicado a los datos de entrenamiento sobre los que se ha creado. Cuando el algoritmo se aplica al conjunto de validación, esa cifra cae ligeramente a un 81 %. Las cifras de los distintos tipos de errores cometidos por el algoritmo se muestran en la tabla 6.1 (esto se denomina matriz de errores, o, a veces, matriz de confusión). Si intentamos detectar supervivientes, el porcentaje de verdaderos supervivientes que son predichos correctamente se conoce como la sensibilidad del algoritmo, mientras que el porcentaje de no supervivientes verdaderos que son correctamente predichos se conoce como la especificidad. Estos términos se derivan de los test de diagnóstico médico.

Aunque es fácil expresar la precisión general, es una medida muy cruda del desempeño, que no tiene en cuenta la fiabilidad de la predicción. Si nos fijamos en las puntas de las ramas del árbol de clasificación, podemos ver que la discriminación que realizan los datos de entrenamiento no es perfecta, y, por ello, en todas las ramas hay algunas personas que sobreviven y otras que no. La regla de asignación cruda simplemente escoge el resultado de la mayoría, pero en lugar de ello podríamos asignar a los nuevos casos una probabilidad de sobrevivir correspondiente a la proporción en el conjunto de entrenamiento. Por ejemplo, alguien con el título «Mr» podría tener asignada una probabilidad de supervivencia del 16 %, en lugar de una predicción categórica simple de que no sobrevivirá.

	Conjunto de entrenamiento			Conjunto de validación		
	No supervivencia predicha	Supervivencia predicha		No supervivencia predicha	Supervivencia predicha	
No sobrevivió	475	93	568	228	45	273
Sobrevivió	71	258	329	35	104	139
	546	351	897	263	149	412
	Fiabilidad = (475 + 258) / 897 = 82 %			**Fiabilidad** = (228 + 104) / 412 = 81 %		
	Sensibilidad = 258 / 329 = 78 %			**Sensibilidad** = 104 / 139 = 75 %		
	Especificidad = 475 / 568 = 84 %			**Especificidad** = 228 / 273 = 84 %		

Tabla 6.1. Matriz de errores de un árbol de clasificación sobre datos de entrenamiento y validación, con información sobre fiabilidad (% de correctamente clasificados), sensibilidad (% de supervivientes correctamente clasificados) y especificidad (% de no supervivientes correctamente clasificados).

Los algoritmos que dan una probabilidad (o cualquier número), más que una simple clasificación, son comparados normalmente con curvas de característica operativa del receptor (ROC)[93] que fueron desarrolladas originalmente en la Segunda Guerra Mundial para analizar las señales de radar. Su mayor virtud es que permiten variar el umbral a partir del cual se predice que la gente sobrevivirá. La tabla 6.1 muestra el efecto de usar un umbral del 50 % para predecir que alguien es un «superviviente», dadas una especificidad y sensibilidad en el conjunto de entrenamiento de 0,84 y 0,78, respectivamente. Pero podríamos haber demandado una mayor probabilidad para predecir si alguien sobrevive, por ejemplo, un 70 %, en cuyo caso la especificidad y la sensibilidad habrían sido del 0,98 y 0,50 respectivamente —con este umbral más exigente, solo identificamos la mitad de los verdaderos supervivientes, pero hacemos menos afirmaciones falsas de supervivencia—. Considerando todos los umbrales posibles para predecir un superviviente, los posibles valores de especificidad y sensibilidad forman una curva. Nótese que un eje de especificidad disminuye convencionalmente de 1 a 0 cuando se traza una curva ROC.

La figura 6.4 muestra las curvas ROC para los conjuntos de entrenamiento y validación. Un algoritmo completamente inútil que

asigne números aleatoriamente tendría una curva ROC diagonal, mientras que los mejores algoritmos tendrían curvas ROC que se moverían hacia la esquina superior izquierda. Una forma estándar de comparar curvas ROC es medir el área que hay por debajo de ellas, hasta la línea horizontal; eso supondría un valor de 0,5 para un algoritmo inútil, y de 1 para uno perfecto que acierte completamente. Para los datos de nuestro conjunto de validación sobre el Titanic, el área bajo la curva ROC es 0,82. Y resulta que hay una interpretación elegante de esa área: si escogemos de manera aleatoria a alguien que sobrevivió realmente y a alguien que no, hay una probabilidad del 82 % de que el algoritmo le asigne al superviviente real una mayor probabilidad de sobrevivir que al no superviviente. Las áreas por encima de 0,8 representan una muy buena capacidad de discriminación.

El área por debajo de la curva ROC sirve para medir lo bien que el algoritmo diferencia a los supervivientes de los no supervivientes, pero no la bondad de las probabilidades. Las personas que están más familiarizadas con las predicciones probabilísticas son los meteorólogos.

Supongamos que queremos predecir si lloverá o no mañana en un momento dado y en un sitio en particular. Unos algoritmos básicos nos darían simplemente una respuesta de sí o no, que podría terminar siendo correcta o errónea. Modelos más sofisticados podrían generar una probabilidad de que llueva, lo que permitiría hacer juicios más precisos —qué hacer si el algoritmo dice que hay una probabilidad del 50 % de que llueva podría ser muy diferente de si dice que esa probabilidad es del 5 %—.

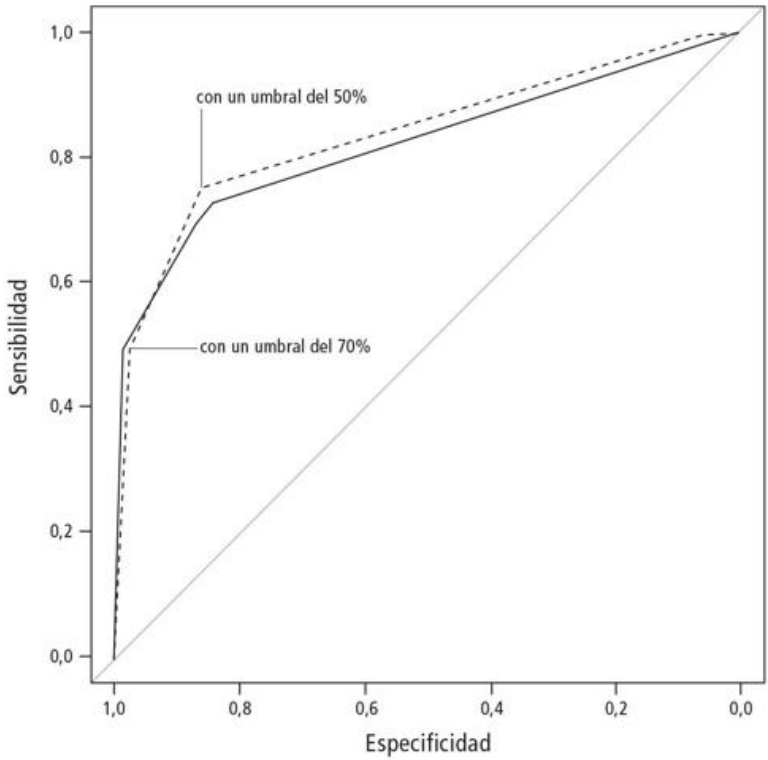


Figura 6.4. Curvas ROC para el árbol de clasificación de la figura 6.3 aplicadas a los conjuntos de entrenamiento (línea discontinua) y de validación (línea sólida). La «sensibilidad» es la proporción de supervivientes correctamente identificados. La «especificidad» es la proporción de no supervivientes correctamente etiquetados como no supervivientes. Las áreas bajo las curvas son de 0,84 y 0,82 para los conjuntos de entrenamiento y validación respectivamente.

¿Cómo sabemos lo buenas que son las predicciones sobre la «probabilidad de precipitación»?

En la práctica, la meteorología se basa en modelos informáticos extremadamente complejos basados a su vez en detalladas fórmulas matemáticas que representan cómo evoluciona el tiempo a partir de las condiciones actuales, y cada estimación del modelo produce una predicción determinista de lluvia del tipo sí o no para un momento y lugar determinados. De manera que para generar una predicción probabilística, el modelo tiene que ser estimado muchas veces, comenzando por unas condiciones iniciales ligeramente ajustadas, que producen una lista de distintos «futuros posibles», en algunos de los cuales llueve, mientras que en otros no. Los meteorólogos estiman un

conjunto de, digamos, cincuenta modelos, y si llueve en un lugar y momento determinados en cinco de esos futuros posibles, afirman que hay una «probabilidad de precipitación» del 10 %.

Pero ¿cómo comprobamos hasta qué punto esas probabilidades son acertadas? No podemos crear una sencilla matriz de errores como en el árbol de clasificación, dado que el algoritmo en ningún momento declara categóricamente si lloverá o no. Podemos crear curvas ROC, pero estas solo se referirán a si en los días en los que llueve se obtienen predicciones más altas que en los que no. Lo más importante es que necesitamos además realizar una calibración, en el sentido de que si juntamos todos los días en los que el meteorólogo dice que hay una probabilidad de lluvia del 70 %, debería llover en aproximadamente un 70 % de esos días. Esto es algo que los meteorólogos se toman muy en serio: las probabilidades deben realmente querer decir lo que dicen, y no sobreestimar o subestimar.

Los gráficos de calibración nos permiten evaluar la fiabilidad de las probabilidades estimadas, juntando, por ejemplo, los eventos a los que se les asigna una probabilidad determinada de ocurrencia, y calculando la proporción de esos eventos que realmente se produjeron.

La figura 6.5 muestra el gráfico de calibración para el árbol de clasificación simple aplicado al conjunto de validación. Queremos que los puntos estén cerca de la línea diagonal, dado que es ahí donde las probabilidades predichas coinciden con los porcentajes observados. Las barras verticales señalan una región en la cual, dadas unas probabilidades predichas fiables, esperaríamos que el porcentaje real se diese en el 95 % de los casos. Si incluimos una línea diagonal, como en la figura 6.5, podemos considerar que nuestro algoritmo está bien calibrado.

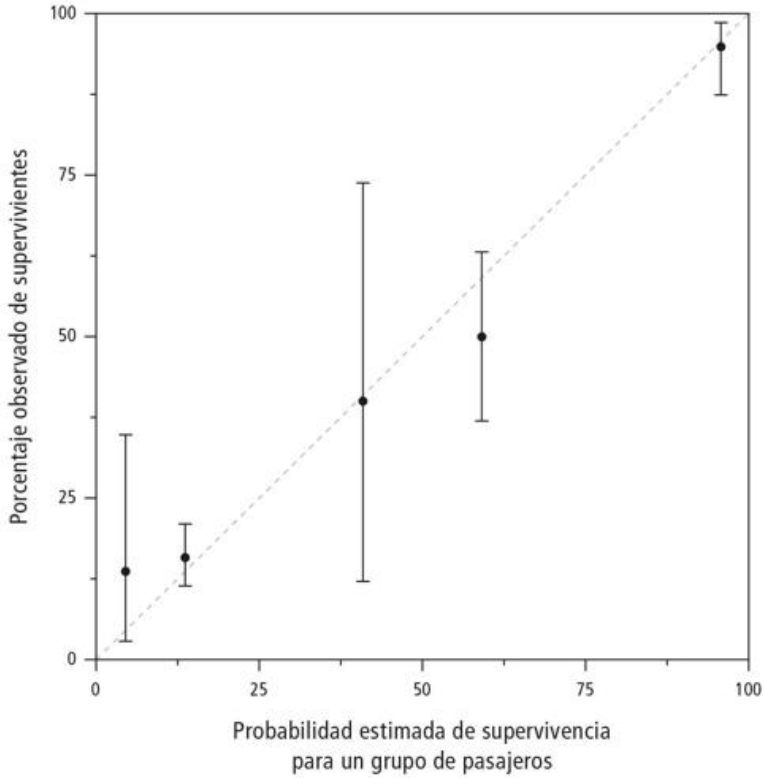


Figura 6.5. Gráfico de calibración para el árbol de clasificación simple que proporciona probabilidades de supervivencia al hundimiento del Titanic, con el porcentaje observado de supervivientes en el eje y, y el porcentaje predicho en el eje x. Queremos que los puntos se encuentren en la línea diagonal, lo que significaría que las probabilidades son fiables y quieren decir lo que dicen.

Una medida combinada de «precisión» para las probabilidades

La curva ROC evalúa lo bien que el algoritmo divide los grupos, y el gráfico de calibración evalúa si las probabilidades quieren decir lo que dicen. Resultaría mejor encontrar una medida compuesta que combine ambos aspectos en una cifra única que pudiésemos utilizar para comparar algoritmos. Afortunadamente, los meteorólogos se ocuparon de ello en la década de 1950.

Si estuviésemos prediciendo una cantidad numérica, como la temperatura que hará mañana a mediodía en un sitio en particular, la fiabilidad normalmente estaría resumida en el error —la diferencia entre la temperatura observada y la predicha—. El resumen normal de los errores a lo largo de varios días es el error cuadrático medio (ECM)

—la media del cuadrado de los errores, algo parecido al criterio de mínimos cuadrados que se usa en el análisis de regresión—.

El truco para calcular probabilidades es usar el mismo error cuadrático medio que cuando predecimos una cantidad, pero asignando un valor 1 a la observación futura de «lluvia» y de 0 a la de «no lluvia». La tabla 6.2 muestra cómo funcionaría este criterio en el caso de un sistema meteorológico ficticio. El lunes se asigna una probabilidad de 0,1 a que llueva, pero resulta que no llueve (la respuesta verdadera es 0), por lo que el error es $0 - 0{,}1 = -0{,}1$. Esto se eleva al cuadrado para obtener 0,01, y así sucesivamente a lo largo de la semana. La media de estos errores al cuadrado, $B = 0{,}11$, es una medida de la precisión (o de falta de ella) del meteorólogo.[94] El error cuadrático medio se conoce como puntuación de Brier, por el meteorólogo Glenn Brier, que describió el método en 1950.

Por desgracia, la puntuación de Brier no es fácil de interpretar en sus propios términos, y, por ello, es difícil saber si un meteorólogo lo está haciendo bien o mal; es mejor comparar su predicción con una puntuación de referencia derivada del historial climatológico. Estas predicciones «basadas en el clima» no tienen en cuenta las condiciones actuales y establecen simplemente la probabilidad de precipitación como el porcentaje de veces en la historia climatológica en las que llovió ese día. Cualquiera puede hacer una predicción de ese tipo sin necesidad de tener ningún conocimiento; en la tabla 6.2 asumimos que esto significa asignar una probabilidad del 20 % de lluvia para cada día de la semana. Esto da una puntuación de Brier para la climatología (lo que llamamos BC) de 0,28.

	Lunes	Martes	Miércoles	Jueves	Viernes	Error cuadrático medio (puntuación de Brier)
«Probabilidad de precipitación»	0,1	0,2	0,5	0,6	0,3	
¿Llovió realmente?	No	No	Sí	Sí	No	
Respuesta verdadera	0	0	1	1	0	
Error	−0,1	−0,2	0,5	0,4	−0,3	
Error al cuadrado	0,01	0,04	0,25	0,16	0,09	B = 0,54 / 5 = 0,11
Probabilidad a partir de datos climatológicos	0,2	0,2	0,2	0,2	0,2	
Error climatológico	−0,2	−0,2	0,8	0,8	0,2	
Error climatológico al cuadrado	0,04	0,04	0,64	0,64	0,04	BC = 1,4 / 5 = 0,28

Tabla 6.2. Predicción ficticia de la «probabilidad de precipitación» de si lloverá o no a mediodía del día siguiente en una localidad específica, con resultado observado: 1 = llovió, 0 = no llovió. El «error» es la diferencia entre el resultado predicho y observado, y el error cuadrático medio es la puntuación de Brier (B). La puntuación climatológica de Brier (BC) se basa en el uso de los porcentajes medios a largo plazo de lluvia en esa época del año como predicciones probabilísticas, que en este caso se asume que son de un 20 % todos los días.

Cualquier algoritmo predictivo decente debería funcionar mejor que las predicciones basadas exclusivamente en el historial climatológico, y, en ese sentido, nuestro sistema predictivo mejora la puntuación en BC − B = 0,28 − 0,11 = 0,17. Los meteorólogos crearon a continuación un «índice de habilidad», que mide la reducción proporcional de la puntuación de referencia: en nuestro caso es 0,61, [95] lo que significa que nuestro algoritmo mejora en un 61 % a un meteorólogo aficionado que use solo datos climatológicos.

Claramente nuestro objetivo es el 100 %, pero solo alcanzaríamos esa cifra si nuestra puntuación de Brier observada se redujese a 0, lo que solo ocurre si predecimos exactamente si lloverá o no. Esto sería mucho esperar de cualquier meteorólogo, y, de hecho, el índice de habilidad para la predicción de lluvia está ahora en torno al 0,4 para

el día siguiente, y al 0,2 para las predicciones a una semana.[96] Por supuesto, la predicción menos trabajada de todas diría simplemente que lo que pase hoy también ocurrirá mañana, lo que proporciona un ajuste perfecto a los datos históricos (hoy), pero podría no ser particularmente buena para predecir el futuro.

Volviendo al desafío del Titanic, considérese el algoritmo sencillo consistente en asignar a todo el mundo una probabilidad del 39 % de supervivencia, que es el porcentaje global de supervivientes en el conjunto de entrenamiento. No estaríamos usando datos individuales, y, en esencia, sería el equivalente a predecir el tiempo con el historial climatológico en lugar de información sobre las circunstancias actuales. La puntuación de Brier para este índice de «ausencia de habilidades» es de 0,232.

Por el contrario, la puntuación de Brier para el árbol de clasificación simple es de 0,139, lo que supone una reducción de un 40 % con respecto a esa predicción de aficionado, y, en ese sentido, demuestra una habilidad considerable. Otra forma de interpretar esta puntuación de Brier de 0,139 es que es exactamente lo que habríamos obtenido si otorgásemos a todos los supervivientes una probabilidad del 63 % de sobrevivir, y a todos los no supervivientes una probabilidad del 63 % de no sobrevivir.

Veremos si podemos mejorar esa puntuación con algunos modelos más complicados, pero primero tenemos que advertir de que quizá es mejor que no sean demasiado complicados.

Sobreajuste

No necesitamos quedarnos en el sencillo árbol de clasificación mostrado en la figura 6.3. Podemos hacer el árbol más y más complejo añadiendo nuevas ramas, lo que nos permitiría clasificar correctamente una parte mayor del conjunto de entrenamiento, a medida que identificamos un creciente número de sus características.

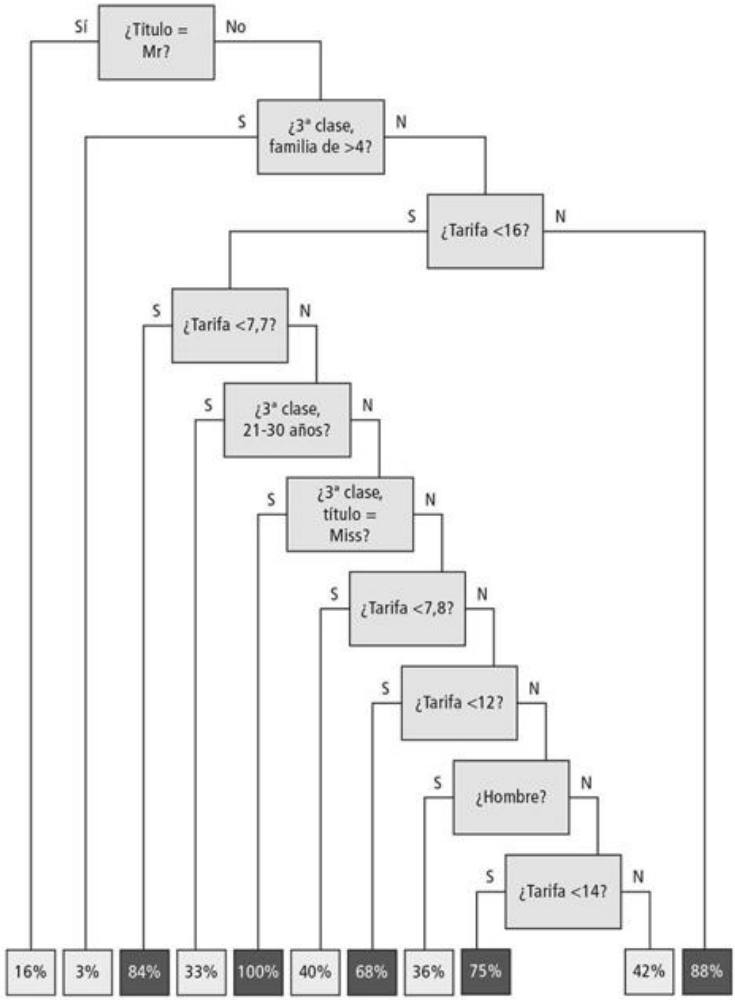


Figura 6.6. Árbol de clasificación sobreajustado para los datos del Titanic. Como en la figura 6.3, el porcentaje al final de cada rama es la proporción de pasajeros en el conjunto de entrenamiento que sobrevivió, y se predice que un nuevo pasajero sobrevivirá si ese porcentaje es mayor del 50 %. El conjunto más bien extraño de preguntas sugiere que el árbol se ha adaptado demasiado a los casos individuales del conjunto de entrenamiento.

La figura 6.6 muestra un árbol de ese tipo, ampliado para poder incluir muchos detalles. Tiene una precisión de 83 % con respecto al conjunto de entrenamiento, mejor que el árbol más pequeño. Pero cuando aplicamos este algoritmo a los datos del conjunto de validación, su precisión se reduce al 81 %, la misma que la del árbol pequeño, y su puntuación de Brier es 0,150, claramente peor que el

árbol sencillo, cuya puntuación es 0,139. Hemos adaptado el árbol a los datos de entrenamiento hasta tal punto que su capacidad predictiva ha comenzado a descender.

Esto se conoce como sobreajuste, y es una de las cuestiones más importantes en la construcción de algoritmos. Al hacer el algoritmo demasiado complejo, esencialmente estamos ajustando el ruido, en lugar de la señal. Randall Munroe (el dibujante conocido por su tira cómica xkcd) ilustró brillantemente lo que supone el sobreajuste. Enunció una serie de «reglas» plausibles seguidas en las elecciones presidenciales de Estados Unidos, solo para ser rotas en elecciones subsiguientes.[97] Por ejemplo:

→ «Ningún republicano ha ganado sin controlar la Cámara o el Senado» (hasta que lo hizo Eisenhower en 1952).

→ «Los católicos no pueden ganar» (hasta Kennedy en 1960).

→ «Nadie ha sido elegido presidente después de un divorcio» (hasta Reagan en 1980).

Y así sucesivamente, incluidas algunas reglas claramente sobrerrefinadas, como la siguiente:

→ «Ningún presidente demócrata sin experiencia en combate ha derrotado a alguien cuyo nombre tenga más puntos en el Scrabble» (hasta que Bill [6 puntos en el Scrabble] Clinton derrotó a Bob [7 puntos en el Scrabble] Dole en 1996).

Sobreajustamos cuando nos adaptamos demasiado a las circunstancias locales, en un esfuerzo meritorio pero equivocado por evitar ser «sesgados» y para tener en cuenta toda la información disponible. Normalmente aplaudiríamos ese deseo de evitar sesgos, pero es un refinamiento que significa que tenemos menos datos con los que trabajar, de manera que la fiabilidad se reduce. El sobreajuste, por tanto, lleva a menos sesgo, pero a costa de una mayor incertidumbre o variación en los estimadores, y es por ello por lo que la protección contra el sobreajuste a menudo se conoce como solución de compromiso sesgo-varianza.

Podemos ilustrar esta sutil idea imaginando una base de datos enorme sobre la vida de la gente que se use para predecir la salud que tendrá usted en el futuro; por ejemplo, su probabilidad de llegar a los 80 años. Podríamos, quizás, observar a todas las personas de su grupo de edad y estatus socioeconómico, y ver qué les pasa; podría haber 10.000 personas con esas características. Si 8.000 de ellas alcanzan los 80 años, podríamos estimar que existe una probabilidad del 80 % de que la gente como usted alcance los 80 años, y estar bastante seguros acerca de esa cifra, dado que se basa en un gran número de personas.

Pero esta estimación solo usa un par de características para emparejarle con los casos en la base de datos, e ignora características más individuales que podrían refinar nuestra predicción: por ejemplo, no hay información sobre su salud actual o sus hábitos. Una estrategia diferente consistiría en encontrar personas que se pareciesen más a usted, con el mismo peso, estatura, presión sanguínea, colesterol, hábitos de ejercicio, consumo de tabaco y alcohol, y así sucesivamente: digamos que le emparejamos con personas con más y más características, hasta reducir la comparación a solo dos personas en la base de datos que serían casi iguales. Supongamos que una de ellas ha alcanzado los 80 años y la otra no. ¿Estimaríamos entonces una probabilidad del 50 % de que usted llegase a cumplir 80? Esa cifra del 50 % es, en este sentido, menos sesgada, dado que la otra persona se parece mucho a usted, pero, dado que solo se basa en dos personas, no es un estimador fiable (a saber, tiene una gran varianza).

Intuitivamente sentimos que hay un punto medio entre estos dos extremos; encontrar ese equilibrio es complicado pero crucial. Una técnica para evitar el sobreajuste es la regularización, que supone estimar modelos complejos, pero donde los efectos de las variables son reducidos a cero. Pero quizá la protección más común es usar la idea sencilla pero potente de la validación cruzada cuando se construye el algoritmo.

Es esencial comprobar cualquier predicción sobre un conjunto de validación independiente que no se haya usado en el entrenamiento del algoritmo, pero eso solo pasa al final del proceso de desarrollo. De manera que, aunque podría mostrar nuestro sobreajuste en ese momento, no nos proporciona un mejor algoritmo. Podemos, no obstante, imitar una validación independiente eliminando, digamos, un 10 % de los datos de entrenamiento, desarrollar el algoritmo sobre el restante 90 %, y la validación sobre ese 10 % eliminado. Esto es la validación cruzada, y puede llevarse a cabo sistemáticamente eliminando el 10 % y repitiendo el procedimiento diez veces, lo que se conoce como validación cruzada decuplicada.

Todos los algoritmos mencionados en este capítulo tienen algunos parámetros regulables cuya función principal es controlar la complejidad del algoritmo final. Por ejemplo, el procedimiento estándar para construir árboles de clasificación es, primero, construir un árbol muy profundo, con muchas ramas, que esté deliberadamente sobreajustado, y después podar el árbol para hacerlo más simple y más robusto: esta poda está controlada por un parámetro de complejidad.

Este parámetro de complejidad puede ser escogido a través del proceso de validación cruzada. Para cada una de las diez muestras de validación cruzada, se desarrolla un árbol para cada uno de los distintos parámetros de complejidad. Para cada valor del parámetro, se calcula el desempeño predicho medio a lo largo de los diez conjuntos de validación cruzada (este desempeño medio tenderá a mejorar hasta cierto punto, y después empeorará a medida que los árboles se hacen demasiado complejos). El valor óptimo del parámetro de complejidad es el que proporciona el mejor desempeño de validación cruzada, y este valor se usa a continuación para construir un árbol a partir del conjunto de entrenamiento completo, que es la versión final.

Hemos empleado la validación cruzada decuplicada para seleccionar el parámetro de complejidad en el árbol de la figura 6.3, y para escoger parámetros ajustados en todos los modelos que consideraremos a continuación.

Modelos de regresión

Vimos en el capítulo 5 que la idea del modelo de regresión es utilizar una fórmula sencilla para predecir un resultado. La variable de respuesta en los datos del Titanic es del tipo sí o no, referida a si se ha sobrevivido o no, y, por ello, lo apropiado es una regresión logística, como en el caso de los datos de cirugía cardiaca infantil de la figura 5.2.

La tabla 6.3 muestra los resultados de una regresión logística. Se ha utilizado el boosting, un procedimiento interactivo diseñado para prestar una mayor atención a los casos más difíciles: a aquellos individuos en el conjunto de entrenamiento que están clasificados incorrectamente en una iteración se les confiere más peso en la siguiente iteración. El número de iteraciones se decide mediante la validación cruzada decuplicada.

Los coeficientes para las características de un pasajero concreto pueden ser sumados para obtener un índice total de supervivencia. Por ejemplo, Francis Somerton empezaría con 3,20, se le restaría 2,30 por ir en tercera clase y 3,86 por recibir el título de «Mr», pero se le sumaría 1,43 por ser un hombre que viaja en tercera clase. Pierde 0,38 por estar en una familia de un único miembro, lo que nos da una puntuación total de −1,91, lo que, a su vez, se traduce en una probabilidad de supervivencia del 13 %, ligeramente inferior al 16 % que nos daría el árbol de clasificación simple.[98]

Se trata de un sistema «lineal», pero nótese que se han incluido interacciones, que son esencialmente características combinadas más complejas, por ejemplo, la puntuación positiva para la interacción de viajar en tercera clase y ser hombre ayuda a contrarrestar las puntuaciones extremadamente negativas de viajar en tercera clase y recibir el título de «Mr», que ya hemos tenido en cuenta. Aunque nos estamos centrando en desempeño predictivo, estos coeficientes sí que proporcionan alguna interpretación de la importancia de las diferentes características.

Características	Puntuación
Puntuación inicial	3,20
Tercera clase	–2,30
«Mr»	–3,86
Hombre en tercera clase	+1,43
Título raro	–2,73
51-60 años en segunda clase	–3,62
Cada miembro de la familia	–0,38

Tabla 6.3. Coeficientes aplicados a características en una regresión logística con datos de supervivientes del Titanic: los coeficientes negativos disminuyen la probabilidad de supervivencia, los coeficientes positivos incrementan esa probabilidad.

Existen análisis de regresión más sofisticados para tratar problemas grandes y complejos, como por ejemplo los modelos no lineales y un proceso conocido como LASSO, que estima simultáneamente coeficientes y selecciona variables independientes relevantes, esencialmente estimando que sus coeficientes sean cero.

Técnicas más complejas

Los árboles de clasificación y los modelos de regresión surgen de filosofías algo distintas sobre cómo modelizar: los árboles intentan construir reglas simples que identifiquen grupos de casos con resultados esperados similares, mientras que los modelos de regresión se centran en el peso que hay que asignar a características específicas, con independencia de cualquier otra cosa que observemos en un caso.

La comunidad dedicada al aprendizaje automático usa árboles de clasificación y regresiones, pero ha desarrollado una amplia gama de métodos alternativos y más complejos para desarrollar algoritmos. Por ejemplo:

→ Bosques aleatorios, que comprenden un gran número de árboles, cada uno de los cuales produce una clasificación. La clasificación final se decide por voto mayoritario, un proceso conocido como empaquetado.

→ Máquinas de vectores de soporte, que intentan encontrar una combinación lineal de aquellas características que separen mejor los distintos resultados.

→ Redes neuronales, que comprenden estratos de nodos, cada uno

de los cuales depende del anterior de manera ponderada, como si se tratase de una serie de regresiones logísticas unas encima de otras. Los pesos de cada estrato se determinan por medio de un procedimiento de optimización, y, como en el caso de los bosques aleatorios, se pueden construir múltiples redes neuronales y extraer una media. Las redes neuronales con muchos estratos se conocen como modelos de aprendizaje profundo: se dice que el sistema de reconocimiento de imágenes de Google, Inception, tiene alrededor de veinte estratos y unos trescientos mil parámetros que estimar.

→ K vecinos más cercanos clasifica de acuerdo con el resultado mayoritario entre casos parecidos del conjunto de entrenamiento.

Los resultados de aplicar algunos de estos métodos a los datos del Titanic, con parámetros ajustados escogidos a través de la validación cruzada decuplicada y ROC como criterio de optimización, se muestran en la tabla 6.4.

La alta precisión de la sumamente elemental regla según la cual «todas las mujeres sobreviven, todos los hombres no sobreviven», que o bien queda por delante o solo un poco por detrás de algoritmos más complejos, demuestra lo inadecuado de una «precisión» cruda como medida de desempeño. El bosque aleatorio produce la mejor discriminación, lo que se refleja en el área por debajo de la curva ROC, aunque, quizás sorprendentemente, las probabilidades provenientes del sencillo árbol de clasificación tienen la mejor puntuación de Brier. No hay, en definitiva, ningún algoritmo claramente ganador. Más adelante, en el capítulo 10, comprobaremos si realmente podemos afirmar con seguridad que hay un ganador en cualquiera de estos criterios, dado que los márgenes de victoria podrían ser tan pequeños como para poder ser explicados como el resultado de una variación aleatoria —como, por ejemplo, quién termina el primero en los conjuntos de validación y entrenamiento—.

Método	Precisión (alta es buena)	Área bajo la curva ROC (alta es buena)	Puntuación de Brier (baja es buena)
Todo el mundo tiene una probabilidad de supervivencia del 39 %	0,639	0,500	0,232
Todas las mujeres sobreviven; todos los hombres no	0,786	0,578	0,214
Árbol de clasificación simple	**0,806**	0,819	**0,139**
Árbol de clasificación (sobreajustado)	**0,806**	0,810	0,150
Regresión logística	0,789	0,824	0,146
Bosque aleatorio	0,799	**0,850**	0,148
Máquina de vectores de soporte (MVS)	0,782	0,825	0,153
Red neuronal	0,794	0,828	0,146
Red neuronal media	0,794	0,837	0,142
K vecinos más cercanos	0,774	0,812	0,180

Tabla 6.3. Coeficientes aplicados a características en una regresión logística con datos de supervivientes del Titanic: los coeficientes negativos disminuyen la probabilidad de supervivencia, los coeficientes positivos incrementan esa probabilidad.

Esto refleja una inquietud general sobre los algoritmos ganadores de competiciones de Kaggle; a saber, que tienden a ser muy complejos, y todo para obtener ese margen minúsculo que se necesita para ganar. Un problema importante es que estos algoritmos tienden a ser cajas negras inescrutables: generan una predicción, pero es casi imposible determinar qué ocurre en su interior. Esto tiene tres consecuencias negativas. En primer lugar, su complejidad extrema hace que su puesta en práctica y su mejora supongan un esfuerzo descomunal: cuando Netflix ofreció un premio de un millón de dólares al mejor sistema predictivo de recomendaciones, el algoritmo ganador era tan complicado que Netflix terminó por no usarlo. Otra característica negativa es que no sabemos cómo se ha alcanzado el resultado, o cuánta confianza podemos tener en este: nuestra única opción es aceptarlo o rechazarlo. Los algoritmos más sencillos se explican mejor. Finalmente, si no sabemos cómo produce sus respuestas un algoritmo, no podemos investigar si tiene un sesgo implícito pero sistemático

contra algunos miembros de la comunidad —algo que desarrollaré más en detalle más adelante—.

Todo ello apunta a que quizá el desempeño cuantitativo no sea el mejor criterio para un algoritmo. Una vez que el desempeño es «lo suficientemente bueno», podría ser razonable preferir que el algoritmo sea sencillo, aun a costa de pequeñas mejoras.

¿Quién fue la persona más afortunada del Titanic?

El superviviente con la puntuación de Brier más alta cuando se hace la media de todos los algoritmos podría ser considerado también como el caso más sorprendente de todos. Se trata de Karl Dahl, un carpintero noruego-australiano de 45 años que viajaba en tercera clase, y que había pagado la misma tarifa que Francis Somerton; dos algoritmos le confirieron una probabilidad de supervivencia del 0 %. Aparentemente se arrojó al agua helada y se subió al bote salvavidas 15, a pesar de que algunos de los pasajeros intentaron impedirlo. Quizá simplemente usó su fuerza.

Esto contrasta con el caso de Francis Somerton, de Ilfracombe, cuya muerte, como hemos visto, encajaba con la pauta general. En lugar de tener un marido exitoso en América, a su mujer, Hannah Somerton, solo le quedaron cinco libras esterlinas, menos de lo que Francis gastó en su billete.

Desafíos planteados por los algoritmos

Los algoritmos pueden mostrar un desempeño notable, pero a medida que su papel en la sociedad aumenta, sus problemas potenciales quedan más al descubierto. Se pueden identificar cuatro de estos problemas.

→ Falta de robustez. Los algoritmos se derivan de asociaciones, y dado que no entendemos cuáles son los procesos subyacentes, pueden ser muy sensibles a los cambios. Aunque solo nos importe la precisión, y no tanto la verdad científica, tenemos que seguir recordando los principios básicos del ciclo PPDAC, y los estadios que hay que seguir, desde los datos obtenidos de una muestra hasta las afirmaciones sobre la población objetivo. Para los análisis predictivos, esta población objetivo incluye casos futuros, y si nada cambia, entonces unos algoritmos construidos sobre datos del pasado deberían funcionar bien. Pero el mundo no siempre permanece igual. Ya hemos mencionado el fracaso de los algoritmos en el cambiante mundo financiero de 2007-2008. Otro ejemplo destacado fue el intento de

Google de predecir las pautas de desarrollo de la gripe a partir de las pautas de búsquedas de los usuarios. Este algoritmo en principio funcionó bien, pero entonces, en 2013, comenzó a sobrepredecir de manera dramática las tasas de contagio de la gripe: una explicación es que los cambios introducidos por Google en el buscador podrían haber generado más términos de búsqueda referidos a la gripe.

→ No tener en cuenta la variabilidad estadística. Las clasificaciones automatizadas basadas en datos limitados son poco fiables. Los docentes en Estados Unidos han sido jerarquizados y penalizados por el desempeño de sus estudiantes en un año concreto, aunque unas clases de menos de treinta alumnos no proporcionan una base fiable para evaluar el valor añadido de un docente. La consecuencia es que los docentes experimentan grandes e implausibles cambios en su evaluación anual: en Virginia, una cuarta parte de los docentes mostraron más de 40 puntos de diferencia en una escala de 1 a 100 de un año a otro.[99]

→ Sesgo implícito. Los algoritmos, como ya hemos dicho, se basan en asociaciones, lo que puede implicar el uso de características que normalmente consideraríamos irrelevantes para las tareas a las que nos enfrentamos. Un algoritmo de visión entrenado para discriminar entre fotos de huskies y de pastores alemanes resultó muy efectivo hasta que dejó de serlo para el caso de huskies que se tienen como mascotas: resultó que su eficacia aparente se basaba en identificar nieve de fondo.[100] Ejemplos menos triviales son el de un algoritmo para identificar la belleza al que no le gustaba la piel oscura, y otro que identificaba a personas negras como gorilas. Aunque algoritmos que pueden tener un gran impacto en la vida de la gente (como es el caso de aquellos que deciden las calificaciones crediticias o de seguros) tuviesen prohibido usar la raza como predictor, podrían usar códigos postales, que son un fuerte proxy de la raza.

→ Falta de transparencia. Algunos algoritmos podrían resultar opacos debido a su gran complejidad. Pero incluso algoritmos sencillos basados en una regresión se volverían opacos si su estructura es privada, quizá por ser un producto comercial patentado. Esta es una de las quejas principales sobre los llamados algoritmos reincidentes, como el caso de la Gestión de Categorización de Delincuentes Penitenciarios para Sanciones Alternativas (COMPAS),[101] de Northpointe, o el Nivel de Inventario de Servicio - Revisado (LSI-R), [102] de MMR.[103] Estos algoritmos producen una puntuación de riesgo o unas categorías que pueden ser utilizadas para guiar las decisiones sobre libertad condicional e imposición de penas, y, a pesar de ello, se desconoce cómo ponderan los distintos factores. Además,

dado que se reúne información sobre los orígenes sociales y las personas con las que ha delinquido anteriormente el individuo, las decisiones no se basan únicamente en el historial delictivo personal, sino en factores de fondo que se ha demostrado que están asociados con un comportamiento delictivo en el futuro, incluso aunque el factor común subyacente sea la pobreza y la privación. Por supuesto, si lo que importa es únicamente lograr una predicción precisa, entonces todo vale, y cualquier factor, incluida la raza, podría ser usado. Pero muchos aducen que la justicia y la imparcialidad exigen que estos algoritmos sean controlados, transparentes y cuyas decisiones se puedan apelar.

Incluso en el caso de los algoritmos patentados, es posible obtener algún tipo de explicación siempre que podamos experimentar con distintos factores empleados para construir el algoritmo. Cuando se adquiere un seguro en Internet, la prima es calculada de acuerdo con una fórmula desconocida sujeta solo a un determinado límite legal: por ejemplo, en el caso de los seguros de automóvil en el Reino Unido, las primas no pueden tener en cuenta el género del cliente, los seguros de vida no pueden usar la raza o información genética excepto la relacionada con la enfermedad de Huntington, y así sucesivamente. Pero podemos hacernos una idea de la influencia de distintos factores mintiendo sistemáticamente y viendo cómo cambia la prima: esto permite cierto grado de ingeniería inversa del algoritmo para ver qué es lo que explica la prima.

Hay una demanda creciente de que los algoritmos que afectan a las vidas de las personas estén bajo control. Asimismo, se está incorporando a la legislación el requisito de que ofrezcan una explicación comprensible de las conclusiones. Estas demandas van en contra de las cajas negras complejas, y podrían llevar a una preferencia por los ciertamente anticuados algoritmos basados en regresiones, en los que la influencia de cada factor está clara.

Pero, tras ver el lado oscuro de los algoritmos, parece apropiado finalizar con un ejemplo que parece completamente beneficioso e ilusionante.

¿Cuál es el beneficio esperado de una terapia complementaria después de cirugía de cáncer de mama?

Casi todas las mujeres recién diagnosticadas de cáncer de mama recibirán algún tipo de cirugía, aunque sea limitada. Una cuestión crítica es la elección de la terapia complementaria que sigue a la cirugía para reducir las posibilidades de recurrencia y de muertes

subsiguientes por cáncer de mama. Las opciones podrían incluir radioterapia, terapia hormonal, quimioterapia y otras opciones de medicamentos. En el contexto del ciclo PPDAC, este es el Problema.

El Plan adoptado por los investigadores del Reino Unido fue desarrollar un algoritmo para ayudarles a tomar esa decisión, basado en 5.700 casos históricos de mujeres con cáncer de mama obtenidos del Registro del Cáncer del Reino Unido. El Análisis comprendía la construcción de un algoritmo que usase información detallada sobre la mujer y su tumor para calcular sus probabilidades de supervivencia hasta diez años después de la cirugía, y cómo cambiaba con distintos tratamientos. Pero hay que tener cuidado en el análisis de los resultados referidos a mujeres que recibieron estos tratamientos en el pasado: recibieron un tratamiento por razones desconocidas y no podemos usar los aparentes beneficios observados en la base de datos. En lugar de ello, se estima un modelo de regresión, con la supervivencia como variable dependiente, pero forzando que el efecto de los tratamientos sea el estimado a partir de análisis de ensayos clínicos a gran escala. El algoritmo subsiguiente está públicamente disponible, y su discriminación y calibración han sido comprobadas mediante bases de datos independientes que comprenden a 27.000 mujeres.[104]

El software informático resultante se denomina Predict 2.1, y los resultados informan (Conclusión y comunicación) del porcentaje de mujeres similares que se espera que sobrevivan cinco y diez años como resultado de distintos tratamientos adicionales. Algunos resultados para una mujer imaginaria se muestran en la tabla 6.5.

Tratamiento	Beneficio adicional con respecto a tratamientos previos	% de supervivencia global
Solo cirugía	–	64 %
+ Terapia hormonal	7 %	70 %
+ Quimioterapia	6 %	76 %
+ Trastuzumab (Herceptin)	3 %	79 %
Para mujeres sin cáncer		87 %

Tabla 6.5. Resultados del algoritmo Predict 2.1: el porcentaje de mujeres de 65 años que se espera que sobrevivan diez años después de la cirugía de cáncer de mama, cuando se detecta un tumor de 2 centímetros de grado 2, dos nodos positivos, y con estatus positivo ER, HER2 y Ki-67. Se muestran los beneficios esperados acumulados para distintos tratamientos complementarios, aunque estos tratamientos podrían tener efectos adversos. El porcentaje de supervivencia para

«mujeres sin cáncer» representa la mejor supervivencia alcanzable, dada la edad de la mujer.

Predict 2.1 no es perfecto, y las cifras de la tabla 6.5 solo pueden emplearse como una guía muy esquemática para un individuo: nos dicen lo que esperaríamos que les pasase a mujeres que son iguales en las características que se incluyen en el algoritmo; para una mujer concreta, deben tenerse en cuenta factores adicionales. No obstante, Predict 2.1 se usa rutinariamente para decenas de miles de casos al mes, tanto en reuniones de equipo interdisciplinares (MTD),[105] en las que se formulan las opciones de tratamiento de un paciente, como en la comunicación de esa información a la mujer. Para aquellas mujeres que desean implicarse plenamente en su tratamiento (un proceso denominado «tratamiento compartido»), puede proporcionar información que normalmente solo está disponible para los terapeutas, y permitirles tener un mayor control sobre sus vidas. El algoritmo no está patentado, el software es libre y el sistema se mejora regularmente para proporcionar más información, incluidos los efectos adversos de los tratamientos.

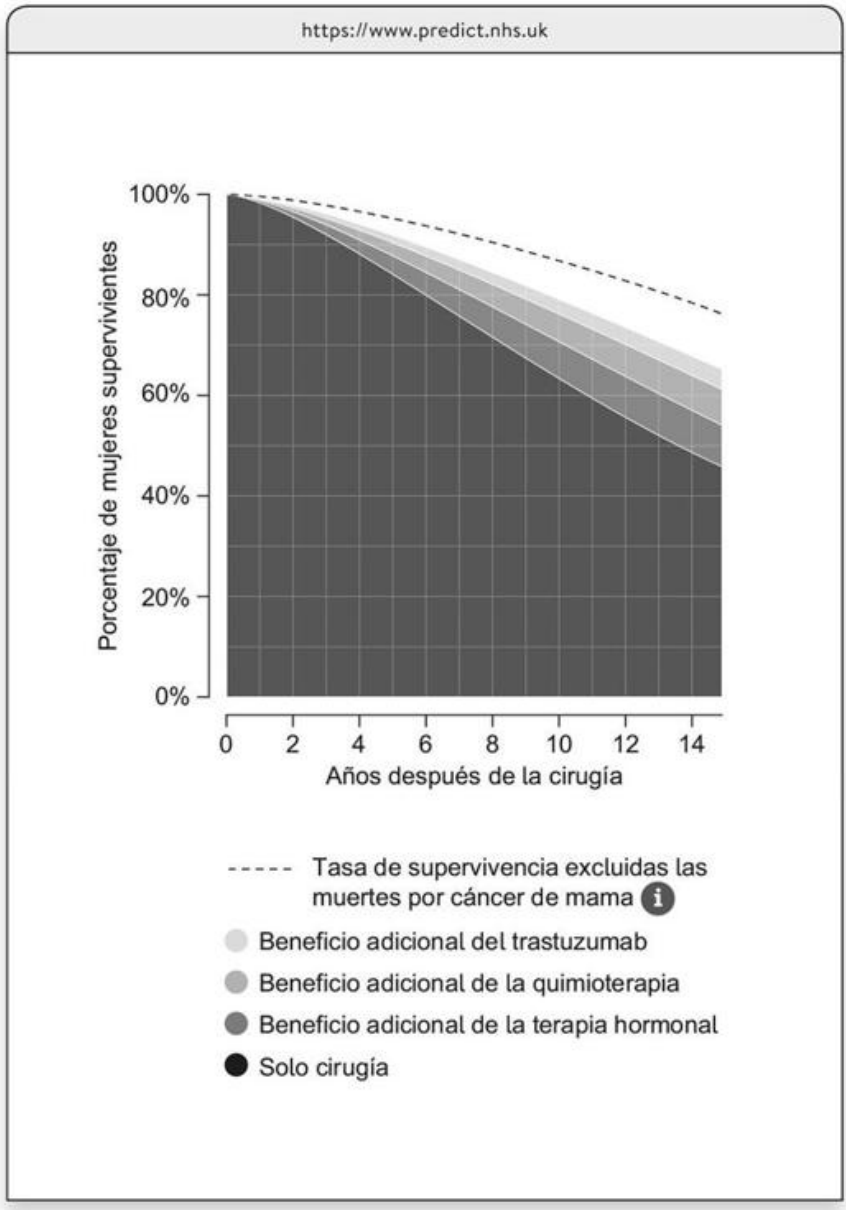


Figura 6.7. Curvas de supervivencia para Predict 2.1 para un máximo de quince años posteriores a la cirugía, para mujeres con las características enumeradas en la leyenda de la tabla 6.5, que muestra la supervivencia adicional acumulada derivada de tratamientos adicionales. El área por encima de la línea discontinua representa a mujeres con cáncer de mama que murieron por otras causas.

Inteligencia artificial

Desde que se utilizó por primera vez en la década de 1950, la idea de la inteligencia artificial (IA) ha experimentado períodos de optimista entusiasmo seguidos de otros más críticos. Yo trabajaba en diagnosis mediante ordenador y manejo de incertidumbre en IA en la década de 1980, cuando gran parte del discurso se planteaba como una competición entre los enfoques basados en probabilidad y estadística, los basados en encapsular «reglas» expertas de juicio y aquellos que intentaban emular las capacidades cognitivas a través de redes neuronales. Desde entonces, este campo ha madurado, con un enfoque más pragmático y universal hacia su filosofía subyacente, aunque el entusiasmo no ha desaparecido.

La IA incluye la inteligencia demostrada por máquinas, que es una idea de amplio alcance. Es un tema mucho mayor que el mucho más restringido de los algoritmos que hemos tratado en este capítulo. El análisis estadístico es, de hecho, solo uno de los componentes del desarrollo de sistemas de IA. Pero, como demuestran los extraordinarios logros recientes de los algoritmos sobre visión, discurso, juegos y demás, el aprendizaje estadístico juega un gran papel en los éxitos de la IA estrecha. Sistemas como Predict, que en otros tiempos habría sido concebido como un sistema de apoyo a las decisiones basado en la estadística, podrían ahora denominarse razonablemente IA.[106]

Muchos de los desafíos enumerados más arriba se reducen a algoritmos que solo modelizan asociaciones, sin la menor idea de los procesos causales subyacentes. Judea Pearl, que en gran medida es responsable de la creciente atención al razonamiento causal en IA, argumenta que estos modelos solo nos permiten responder a preguntas del tipo: «Hemos observado X, ¿qué esperamos observar a continuación?». Mientras, la IA general necesita de un modelo causal para determinar cómo funciona realmente el mundo, lo que permitiría responder a preguntas a nivel humano referidas al efecto de las intervenciones («¿Qué ocurre si hacemos X?») y contrafactuales («¿Qué habría ocurrido si no hubiésemos hecho X?»).

Estamos muy lejos de que la IA tenga esta capacidad.

Este libro enfatiza el clásico problema estadístico de las muestras pequeñas, los sesgos sistemáticos (en el sentido estadístico) y la falta de generalización a nuevas situaciones. La lista de los desafíos a los que se enfrentan los algoritmos muestra que, aunque tener un gran volumen de datos podría reducir la inquietud acerca del tamaño muestral, los otros problemas tienden a empeorar, y nos enfrentamos al problema adicional de explicar el razonamiento de un algoritmo.

Tener toneladas de datos solo incrementa los desafíos que acarrea generar conclusiones sólidas y responsables. Cuando se construyen algoritmos, ser humilde resulta crucial.

Resumen

→ Los algoritmos construidos a partir de datos pueden emplearse para la clasificación y predicción en aplicaciones tecnológicas.

→ Es importante cuidarse de sobreajustar un algoritmo a datos de entrenamiento, lo que es esencialmente ajustar al ruido más que a la señal.

→ Los algoritmos pueden ser evaluados por su precisión clasificatoria, su capacidad de discriminar entre grupos y su precisión predictiva global.

→ Los algoritmos complejos pueden carecer de transparencia, y podría valer la pena aumentar la comprensión a costa de la precisión.

→ El uso de los algoritmos y la inteligencia artificial presenta muchos desafíos, y es vital tener conocimiento tanto sobre el poder como sobre las limitaciones de los métodos de aprendizaje automático.

[90] M. Friendly et al., «HistData: Data Sets from the History of Statistics and Data Visualization», 2018, https://CRAN.r-project.org/package=HistData.

[91] Estos títulos incluyen «Dona», «Lady», «Countess», «Capt», «Col», «Don», «Dr», «Major», «Rev.», «Sir», «Jonkheer».

[92] Para no tener que esperar al final de la competición (en 2020 para los datos del Titanic), antes de que nadie obtenga la información, Kaggle divide el conjunto de validación en conjuntos público y privado. Las puntuaciones de los competidores en el conjunto público son publicadas en una tabla de líderes, y esto proporciona una clasificación provisional accesible a todo el mundo. Pero el desempeño en el conjunto privado es lo que se usa realmente para evaluar la clasificación final de los competidores cuando termina la competición.

[93] Por sus siglas en inglés, Receiver Operating Characteristic. (N. del T.).

[94] Podría ser tentador usar el «error absoluto», lo que querría decir

que perderías 0,1 cuando asignas una probabilidad del 10 % a un evento que no ocurre, frente al error al cuadrado, que sería 0,01. Esta elección aparentemente inocua sería un gran, gran error. Alguna teoría bastante básica muestra que este error «absoluto» llevaría a la gente a exagerar racionalmente su confianza para minimizar su error esperado, y prever una probabilidad de lluvia del «0 %», aunque pensasen genuinamente que la probabilidad es del 10 %.

[95] El índice de habilidad es (BC − B) $BC = 1 - B$ BC = 1 − 0,11 / 0,28 = 0,61.

[96] Probabilidad verificada de precipitaciones: http://www.cawcr.gov.au/projects/verification/POP3/POP3.html.

[97] «Electoral Precedent», xkcd, https://xkcd.com/1122/.

[98] Para transformar una puntuación total S en una probabilidad de supervivencia p, usaríamos la fórmula p = 1 / (1 + e−S), donde e es la constante exponencial. Esto es el inverso de la ecuación de regresión logística loge p / (1 − p) = S.

[99] A partir del libro Armas de destrucción matemática (Capitán Swing, 2017), de Cathy O'Neil, que proporciona muchos ejemplos del mal uso de los algoritmos.

[100] http://innovation.uci.edu/2017/08/husky-or-wolf-using-a-black-box-learningmodel-to-avoid-adoption-errors/.

[101] Por sus siglas en inglés, Correctional Offender Management Profiling for Alternative Sanctions. (N. del T.).

[102] Por sus siglas en inglés, Level of Service Inventory - Revised. (N. del T.).

[103] El uso de los algoritmos COMPAS y MMR es criticado en C. O'Neil, Armas de destrucción matemática.

[104] NHS, Predict: Breast Cancer (2.1): http://www.predict.nhs.uk/predict_v2.1/.

[105] Por sus siglas en inglés, multidisciplinary team meetings. (N. del T.).

[106] Quizá simplemente para atraer financiación…

07

¿Qué seguridad nos merecen los resultados?

Estimadores e intervalos

¿Cuánta gente está desempleada en el Reino Unido?

En enero de 2018, la página web de la BBC anunció que en los tres meses transcurridos desde noviembre «el desempleo en el Reino Unido disminuyó en 3.000 personas, hasta los 1,44 millones». Las razones de esta caída son debatibles, pero nadie cuestionó si la cifra era realmente fiable. Sin embargo, un escrutinio cuidadoso de la página web de la Oficina de Estadísticas Nacionales del Reino Unido reveló que el margen de error de estas cifras era de ±77.000; en otras palabras, el cambio real podría haber estado entre una caída de 80.000 y un aumento de 74.000. De manera que, aunque los periodistas y los políticos parecían creer que este supuesto descenso de 3.000 era una cantidad fija e inmutable para todo el país, se trataba de hecho de un estimador impreciso basado en una encuesta a unas 100.000 personas.[107] Igualmente, cuando la Oficina de Estadísticas Laborales de Estados Unidos informó de un aumento desestacionalizado del empleo civil de 108.000 desde diciembre de 2017 a enero de 2018,[108] se basaba en una muestra de alrededor de 60.000 familias y tenía un margen de error (de nuevo difícil de encontrar en el informe) de ±300.000.

Reconocer la incertidumbre es importante. Cualquiera puede hacer una estimación, pero ser capaz de evaluar de modo realista su posible error es algo crucial para la ciencia estadística. Incluso aunque implique asumir algunos conceptos complejos.

Supongamos que hemos reunido unos datos precisos, quizá por medio de una encuesta bien diseñada, y queremos generalizar los resultados a nuestra población objeto de estudio. Si hemos sido cuidadosos y hemos evitado sesgos internos, digamos, por ejemplo, mediante una muestra aleatoria, entonces deberíamos esperar que los estadísticos de resumen calculados a partir de la muestra estén cerca de los valores correspondientes de la población objeto de estudio.

Vale la pena que nos detengamos en esta importante cuestión. En un estudio bien construido, esperamos que la media de nuestra muestra esté cerca de la media de la población, que el rango intercuartílico de la muestra esté cerca del rango intercuartílico de la población, y así

sucesivamente. Ya vimos el concepto de resúmenes poblacionales, ilustrado con los datos del peso al nacer del capítulo 3, donde denominamos «estadístico» a la media de la muestra, y «parámetro» a la de la población. Usando una estricta terminología estadística, estas dos cifras se distinguen normalmente asignándoles letras latinas y griegas, respectivamente, en un intento posiblemente destinado al fracaso de evitar confusión; por ejemplo, m representa normalmente la media muestral, mientras que la letra griega μ (mu) es la media de la población; y s generalmente representa la desviación típica de la muestra, mientras que σ (sigma), la de la población.

A menudo solo se presentan los estadísticos de resumen, y esto es suficiente en algunas circunstancias. Por ejemplo, hemos visto que la mayoría de la población no es consciente de que las cifras de desempleo para el Reino Unido y Estados Unidos no se basan en un recuento completo de aquellas personas oficialmente registradas como desempleadas, sino más bien en grandes encuestas. Si una encuesta de ese tipo encuentra que el 7 % de la muestra está desempleada, las agencias nacionales y los medios de comunicación normalmente presentan este valor como si el 7 % de la población estuviese desempleada, en lugar de reconocer que ese 7 % es solo una estimación. En términos más precisos, confunden la media de la muestra con la de la población.

Esto no importaría mucho si simplemente quisiéramos ofrecer una imagen general de lo que ocurre en el país, y si la encuesta es enorme y fiable. Pero supongamos, por poner un ejemplo extremo, que nos enteramos de que solo se ha preguntado a 100 personas si estaban desempleadas, y 7 de ellas dicen que lo están. La estimación sería un 7 %, pero probablemente no la consideraríamos muy fiable, y no estaríamos muy satisfechos si ese valor fuese tratado como una descripción de toda la población. ¿Y si el tamaño de la encuesta fuese de 1.000 personas? ¿De 100.000? Con una encuesta suficientemente grande, podríamos empezar a sentirnos más cómodos con la idea de que una estimación a partir de una muestra es un buen resumen de lo que pasa. El tamaño muestral debería afectar a nuestra confianza en el estimador, y saber hasta qué punto ese tamaño supone una diferencia en los estimadores es una necesidad básica para poder realizar una inferencia estadística correcta.

Número de parejas sexuales

Volvamos a la encuesta Natsal que vimos en el capítulo 2, en la que se preguntaba a los participantes cuántas parejas sexuales habían tenido a lo largo de su vida. En el rango de edades de 35-44 había 1.125

mujeres y 806 hombres, por lo que podemos decir que los estadísticos muestrales de resumen de la tabla 2.2 se habían calculado a partir de una gran encuesta. Estos indicaban una mediana de 8 parejas declaradas para los hombres y 5 para las mujeres. Dado que sabemos que la encuesta se basaba en una adecuada muestra aleatoria, es razonable asumir que la población objeto de estudio se asemejaba a la población objetivo, que en este caso es la población británica adulta. La cuestión crucial es: ¿cuánto se acercan estas estadísticas a lo que habríamos encontrado si hubiésemos podido preguntar a todo el mundo?

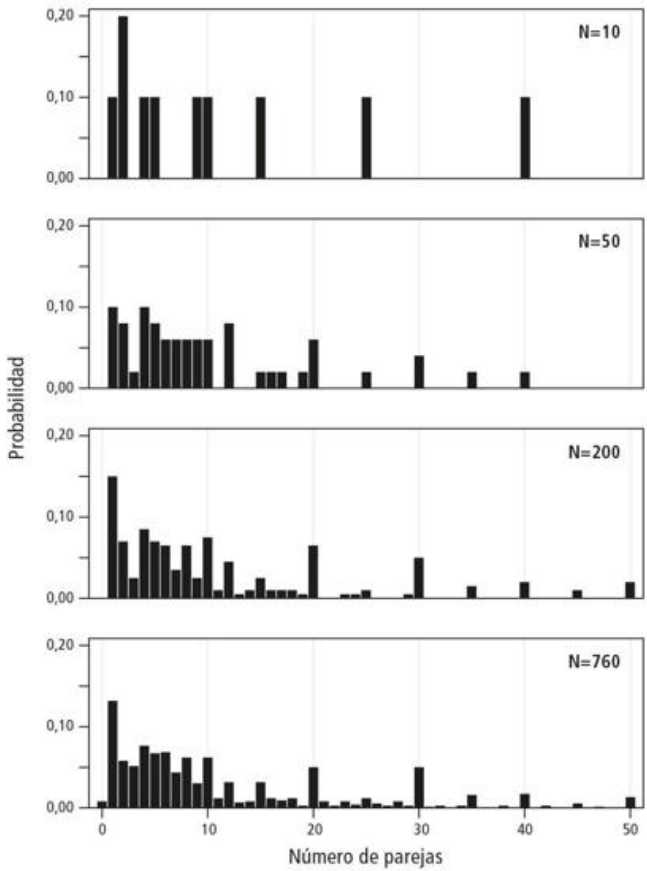


Figura 7.1. El cuadro inferior refleja la distribución de respuestas para los 760 hombres de la encuesta. Los individuos son sucesivamente seleccionados en muestras aleatorias a partir de este grupo, con un tamaño de 10, 50, 200, y generando las distribuciones que se expresan en los tres cuadros superiores. Los tamaños muestrales más pequeños indican una pauta más variable, pero la forma de la distribución gradualmente se aproxima a la de todo el grupo de 760 hombres. No

se muestran valores por encima de 50 parejas.

Para ilustrar cómo depende la precisión de los estadísticos del tamaño muestral, vamos a pretender por el momento que los hombres encuestados representan de hecho la población que nos interesa. El cuadro inferior de la figura 7.1 muestra la distribución de los 760 hombres que declararon haber tenido hasta 50 parejas. Después tomamos muestras sucesivas de individuos de esta «población» de 760 hombres, deteniéndonos cuando alcanzamos 10, 50 y 200 hombres. La distribución de los datos de estas muestras se recoge en la figura 7.1 —está claro que las muestras más pequeñas son más irregulares, por ser más sensibles a puntos de datos específicos—. Los estadísticos de resumen para las muestras sucesivas, más grandes, se muestran en la tabla 7.1, en la que se aprecia que el número bastante bajo de parejas (una media de 8,3) en la primera muestra de diez individuos es sistemáticamente sobrepasado, a medida que el tamaño de la muestra aumenta y los estadísticos se acercan más y más a los del grupo completo de 760 hombres.

Tamaño muestral	Número medio de parejas	Número mediano de parejas
10	8,3	9
50	10,5	7,5
200	12,2	8
760	11,4	7

Tabla 7.1. Estadísticos de resumen para el número de parejas sexuales a lo largo de la vida declaradas por hombres de 35 a 44 años en Natsal-3, para muestras sucesivas cada vez mayores y para los datos completos de 760 hombres.

Volvamos ahora al problema que nos ocupa: ¿qué podemos decir sobre la media y la mediana de parejas en toda la población objeto de estudio de hombres entre 35 y 44 años, basándonos en las muestras reales de hombres que aparecen en la figura 7.1? Podríamos estimar estos parámetros poblacionales a partir de los estadísticos muestrales de cada grupo de la tabla 7.1, presumiendo que aquellos basados en las muestras mayores son en cierta manera «mejores»: por ejemplo, los estimadores del número medio de parejas convergen hacia 11,4, y con una muestra suficientemente amplia podríamos presumiblemente acercarnos tanto como queramos a la respuesta real.

Ahora llegamos a un paso crítico. Para determinar cómo son de precisos estos estadísticos, necesitamos saber cuánto cambiarían si

nosotros (en nuestra imaginación) repitiésemos el proceso muestral muchas veces. En otras palabras, si extrajésemos repetidamente muestras de 760 hombres del país, ¿cuánto cambiarían los estadísticos que hemos calculado?

Si supiésemos cuánto variarían estos estimadores, eso nos ayudaría a determinar la precisión de nuestro estimador real. Pero, desgraciadamente, solo podemos determinar la variabilidad precisa de nuestros estimadores si sabemos los detalles de la población. Y eso es exactamente lo que no sabemos.

Hay dos formas de resolver esta circularidad. La primera es hacer algunas asunciones matemáticas sobre la forma de la distribución poblacional, y usar teoría probabilística sofisticada para determinar la variabilidad que podríamos esperar en nuestra estimación, y, por consiguiente, lo lejos que podríamos esperar que esté, por ejemplo, la media de nuestra muestra de la media de la población. Este es el método tradicional que se enseña en los libros de texto de Estadística, y veremos cómo funciona en el capítulo 9.

No obstante, hay un enfoque alternativo, basado en la asunción plausible de que la población debería parecerse a la muestra. Dado que no podemos extraer repetidamente nuevas muestras de la población, ¡lo que hacemos es extraer repetidamente nuevas muestras de nuestra muestra!

Podemos ilustrar esta idea con nuestra muestra previa de 50, recogida en el cuadro superior de la figura 7.2, que tiene una media de 10,5. Supongamos que extraemos 50 puntos de datos de manera secuencial, reemplazando en cada caso el dato que hemos extraído, y obtenemos la distribución de datos del segundo cuadro, que tiene una media de 8,4.[109]

Nótese que esta distribución puede contener únicamente puntos de datos que adoptan los mismos valores que la muestra original, pero contendrán diferentes números de cada valor y, por ello, la forma de la distribución será ligeramente diferente, con una media ligeramente distinta. Este proceso puede repetirse, y, de hecho, la figura 7.2 muestra tres re-muestreos de ese tipo, con medias de 8,4, 9,7 y 9,8.

Por tanto, nos podemos hacer una idea de cómo varía nuestro estimador a través de este ejemplo de re-muestreo con reemplazo. Esto se conoce como bootstrap: la idea mágica de elevarse a uno mismo por el propio bootstrap se refleja en esta capacidad de aprender acerca de la variabilidad de un estimador sin hacer ninguna asunción sobre la

forma de la distribución de la población.

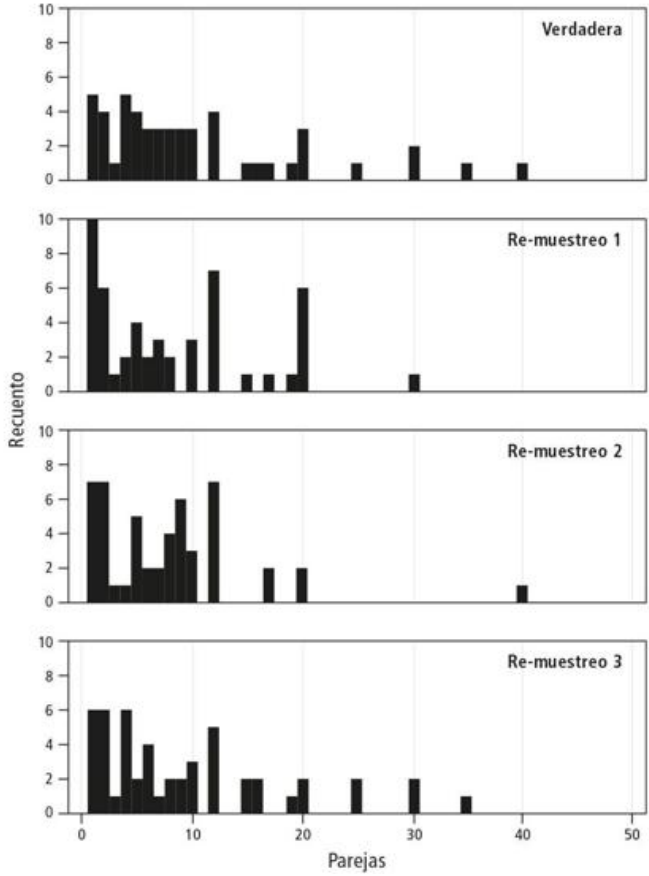


Figura 7.2. Muestra original de 50 observaciones, y tres re-muestreos bootstrap, cada uno de los cuales está basado en un muestreo de 50 observaciones aleatorias a partir del conjunto original, reemplazando cada vez los puntos de datos muestrales. Por ejemplo, una observación de 25 parejas ocurre una vez en los datos originales. Este punto de datos no está incluido en la primera o segunda muestra bootstrap, pero aparece dos veces en la tercera.

Si repetimos este re-muestreo, digamos, 1.000 veces, obtenemos 1.000 estimadores posibles de la media. Se presentan como histogramas en el segundo cuadro de la figura 7.3. Los otros cuadros reflejan los resultados de hacer bootstrap con las otras muestras de la figura 7.1; cada histograma expresa la dispersión de los estimadores bootstrap alrededor de la media de la muestra original. Esto se conoce como distribuciones muestrales de los estimadores, dado que reflejan la variabilidad en estimadores que surgen de un muestreo repetido de los datos.

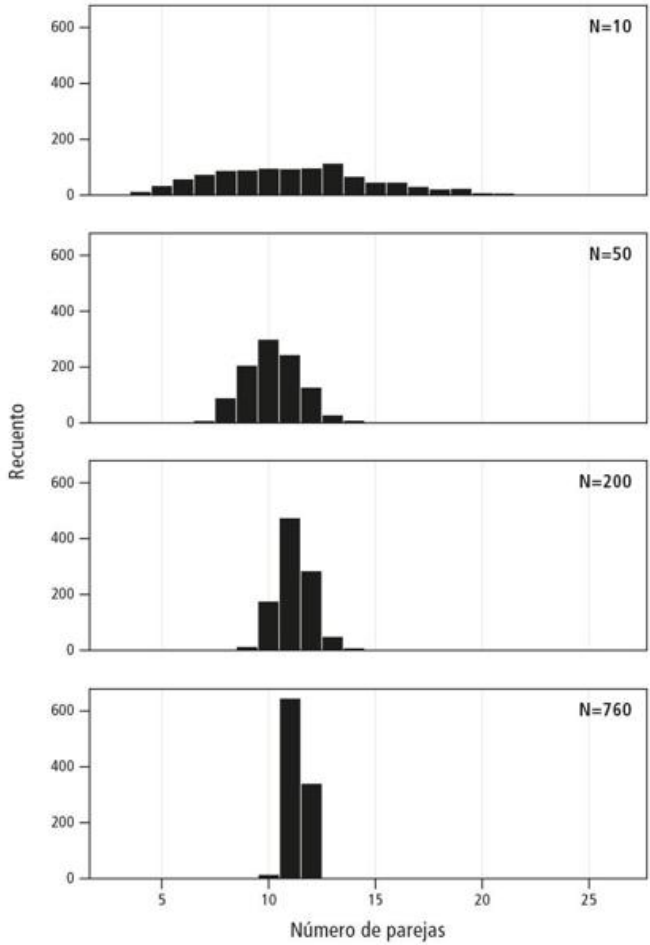


Figura 7.3. Distribución de medias muestrales de 1.000 re-muestreos bootstrap, para cada una de las muestras originales de tamaño 10, 50, 200 y 760 de la figura 7.1. La variabilidad de las medias muestrales de los re-muestreos bootstrap disminuye a medida que se incrementa el tamaño de la muestra.

La figura 7.3 ilustra algunas características claras. La primera, y quizás la más notable, es que casi toda traza de asimetría que veíamos en las muestras originales ha desaparecido —las distribuciones de los estimadores basados en los datos re-muestreados son casi simétricas alrededor de la media de los datos originales—. Este es un primer acercamiento a lo que se conoce como teorema central del límite, que dice que la distribución de la media de una muestra tiende hacia la forma de una distribución normal a medida que se incremente el tamaño de la muestra, casi con independencia de la forma de la distribución original de los datos. Es este un resultado excepcional,

que exploraremos con más detalle en el capítulo 9.

Tamaño de la muestra	Número medio de parejas	Intervalo de incertidumbre *bootstrap* del 95 %
10	8,3	5,3 a 11,5
50	10,5	7,7 a 13,8
200	12,2	10,5 a 13,8
760	11,4	10,5 a 12,2

Tabla 7.2. Medias muestrales para el número de parejas sexuales a lo largo de la vida declaradas por hombres de 35 a 44 años, para muestras aleatorias anidadas de tamaño 10, 50, 200 y datos completos para 760 hombres, con intervalos de incertidumbre bootstrap del 95 %, también conocidos como márgenes de error.

Algo muy importante es que estas distribuciones bootstrap nos permiten cuantificar nuestra incertidumbre acerca de los estimadores que se reflejan en la tabla 7.1. Por ejemplo, podemos encontrar el rango de valores que contiene el 95 % central de las medias de los re-muestreos bootstrap, y denominarlo como intervalo de incertidumbre del 95 % para los estimadores originales, o, alternativamente, margen de error. Todo ello se expresa en la tabla 7.2 —la simetría de las distribuciones bootstrap significa que los intervalos de incertidumbre son aproximadamente simétricos alrededor del estimador original—.

La segunda característica importante de la figura 7.3 es que las distribuciones bootstrap se hacen más estrechas a medida que se incrementa el tamaño de la muestra, lo que se refleja en los cada vez más pequeños intervalos de incertidumbre al 95 %.

Esta sección ha introducido algunas ideas difíciles pero importantes:

→ La variabilidad en la estadística basada en muestras.

→ El uso del bootstrap cuando no queremos hacer asunciones sobre la forma de la población.

→ El hecho de que la forma de la distribución de los estadísticos no depende de la forma de la distribución original de la cual se extraen los puntos de datos.

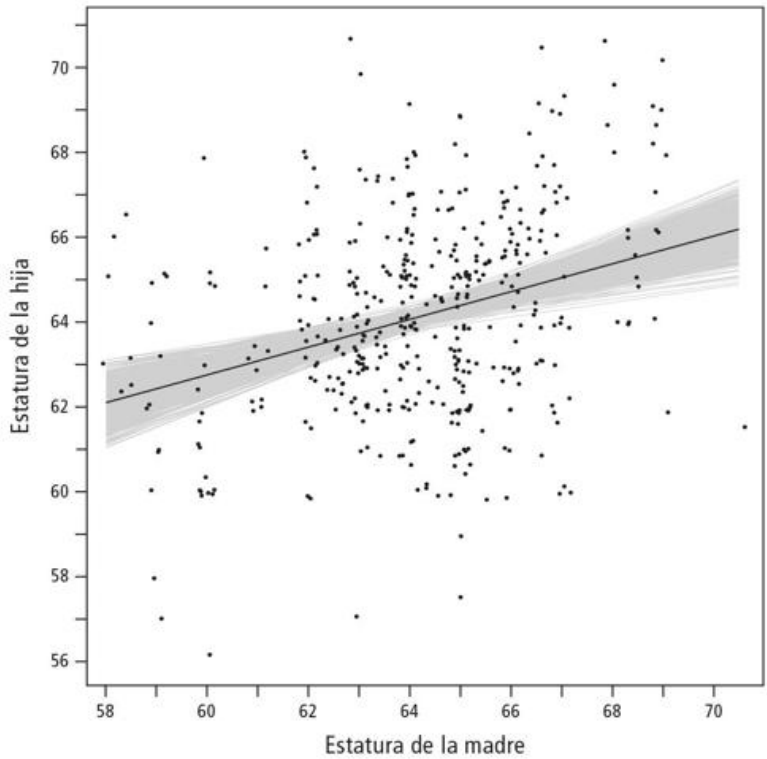


Figura 7.4. Líneas de regresión ajustadas para veinte re-muestreos bootstrap de los datos de estatura madre-hija superpuestos sobre los datos originales, donde se aprecia la relativamente pequeña variabilidad del gradiente como consecuencia del gran tamaño de la muestra.

Es de destacar que hemos logrado explicar todo esto sin recurrir a las matemáticas, excepto con relación a la idea de la extracción aleatoria de observaciones.

Ahora mostraré que la misma estrategia a partir de bootstrap puede ser aplicada a situaciones más complejas.

En el capítulo 5 ajustamos líneas de regresión a los datos de Galton sobre estatura, lo que nos permitió predecir, por ejemplo, la estatura de una hija basándonos en la estatura de su madre, utilizando una línea de regresión con un gradiente estimado de 0,33 (tabla 5.2). Pero ¿cuánta confianza podemos tener acerca de la posición de esa línea que hemos ajustado a los datos? Hacer bootstrap nos proporciona una forma intuitiva de responder a esta pregunta sin hacer asunciones matemáticas sobre la población subyacente.

Para hacer bootstrap con las 433 parejas hija-madre que se muestran en la figura 7.4, se extrae un re-muestreo de 433, con reemplazamiento, a partir de los datos, y se ajusta una línea de mínimos cuadrados («mejor ajuste»). Esto se repite tantas veces como se desee: por poner un ejemplo, la figura 7.4 muestra las líneas ajustadas que se derivan de solo veinte re-muestreos, y se puede percibir claramente la dispersión de las líneas. Queda claro que, dado que los datos originales son grandes, hay una variabilidad relativamente pequeña en las líneas ajustadas, y cuando nos basamos en 1.000 re-muestreos bootstrap, un intervalo del 95 % para el gradiente iría de 0,22 a 0,44.

El bootstrap proporciona una forma intuitiva (y calculable por ordenador) de evaluar la incertidumbre en nuestros estimadores, sin hacer fuertes asunciones y sin usar la teoría de la probabilidad. Pero esa técnica no es factible cuando se trata, por ejemplo, de establecer los márgenes de error de encuestas sobre desempleo de 100.000 personas. Aunque el bootstrap es una idea sencilla, brillante y extraordinariamente efectiva, aplicarla a una cantidad tan grande de datos es simplemente demasiado rudimentario, especialmente cuando existe una teoría conveniente que puede generar fórmulas para calcular la anchura de los intervalos de incertidumbre. Pero antes de demostrar esta teoría en el capítulo 9, debemos primero enfrentarnos a la deliciosa —pero compleja— teoría de la probabilidad.

Resumen

→ Los intervalos de incertidumbre son importantes para la presentación de los resultados estadísticos.

→ Hacer bootstrap sobre una muestra consiste en crear nuevos conjuntos de datos del mismo tamaño re-muestreando los datos originales, con reemplazamiento.

→ Los estadísticos de la muestra calculados a partir de re-muestreos derivados de un bootstrap tienden hacia la distribución normal para conjuntos de datos mayores, con independencia de la forma de la distribución de datos original.

→ Los intervalos de incertidumbre basados en bootstrap sacan partido de la potencia computacional actual, no requieren asunciones sobre la forma matemática de la población y no requieren una teoría de la probabilidad compleja.

[107] Cuando sugerí en una ocasión a un grupo de periodistas que esto debía indicarse claramente en sus artículos, me encontré con una absoluta incomprensión.

[108] Los cambios en el desempleo derivados de datos de nóminas se basan en informaciones proporcionadas por los empleadores y son algo más precisos, con un margen de error de alrededor de ±100.000.

[109] Pensemos en una bolsa de 50 bolas, cada una etiquetada como un punto de datos de nuestra muestra de 50; por ejemplo, una podría ser etiquetada «25», dos serían etiquetadas «30», y así sucesivamente. Extraemos una bola aleatoriamente de la bolsa, registramos su valor y después la reemplazamos, restaurando el número de bolas en la bolsa a 50. Repetimos este proceso de extracción, registro y reemplazo un total de 50 veces, generando una distribución de puntos de datos similares al «re-muestreo 1».

08

Probabilidad: el lenguaje de la incertidumbre y la variabilidad

En la Francia de la década de 1650, el elegante Chevalier [Caballero] de Méré tenía un problema de juego. No era tanto que apostase demasiado (aunque lo hacía), sino que quería saber en cuál de dos juegos tenía la mayor probabilidad de ganar:

Juego 1. Lanzar un dado un máximo de cuatro veces, y ganar si sacas un seis.

Juego 2. Lanzar dos dados un máximo de veinticuatro veces, y ganar si sacas un doble seis.

¿Cuál era la mejor apuesta?

Siguiendo buenos principios estadísticos, el Chevalier de Méré decidió jugar ambos juegos repetidas veces para ver con qué frecuencia ganaba. Esto le llevó mucho tiempo y esfuerzo, mientras que en un extraño universo paralelo en el que hubiese ordenadores pero no teoría de la probabilidad, el bueno del Chevalier (cuyo nombre real era Antoine Gombaud) no habría perdido el tiempo reuniendo esos datos, simplemente habría simulado miles de juegos.

La figura 8.1 refleja los resultados de una simulación de ese tipo, y muestra cómo la proporción global de victorias en cada juego cambia a medida que se «juega» más y más. Aunque el juego 2 parece ser la mejor apuesta durante un tiempo, tras unas 400 rondas de cada uno de ellos queda claro que el juego 1 es mejor, y a (muy) largo plazo, podemos esperar ganar alrededor de un 52 % de los juegos 1, y solo el 49 % de los juegos 2.

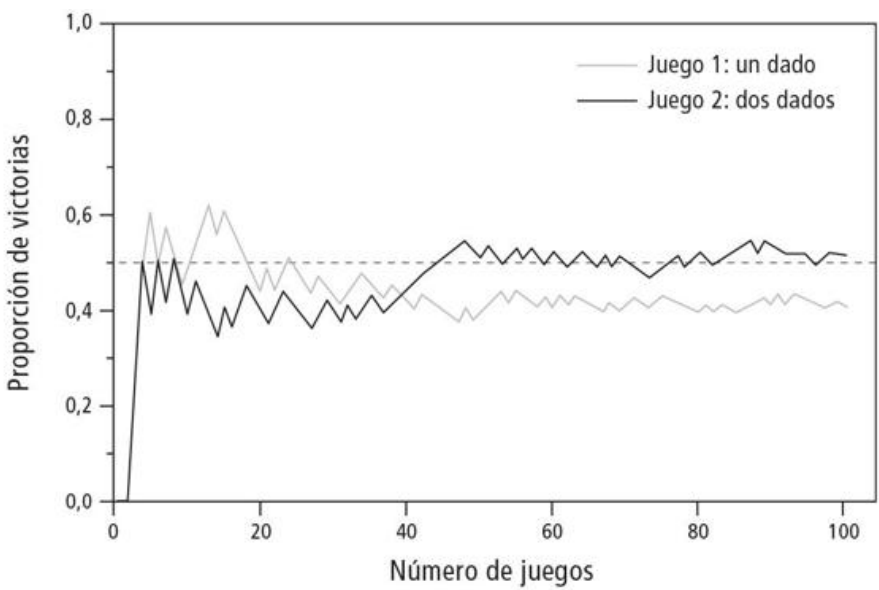
Es de destacar que el Chevalier jugó con tanta frecuencia que llegó a la misma conclusión: el juego 1 era la apuesta marginalmente mejor. Esto iba en contra de sus (erróneos) intentos de calcular las probabilidades de ganar,[110] por lo que buscó ayuda en el elegante salón Mersenne de París. Afortunadamente, el filósofo Blaise Pascal era también miembro del salón, y Pascal, a su vez, escribió a su amigo Pierre de Fermat (el autor del famoso «último teorema») sobre los problemas planteados por el Chevalier. Juntos, dieron los primeros pasos en la teoría de la probabilidad.

A pesar del hecho de que durante milenios los humanos han apostado jugando con huesos o dados, la teoría formal de la probabilidad es una

idea comparativamente reciente. Después del trabajo de Pascal y Fermat en la década de 1650, la base matemática de la teoría fue puesta a punto en los siguientes cincuenta años, y en la actualidad la probabilidad se aplica a la física, los seguros, las pensiones, el comercio financiero, la predicción meteorológica y, por supuesto, las apuestas. Pero ¿por qué necesitamos usar la teoría de la probabilidad para la estadística?

Ya hemos visto el concepto de la «selección aleatoria» de puntos de datos a partir de una distribución poblacional —la amiga con el bebé con poco peso al nacer del capítulo 3 fue nuestra primera introducción a la probabilidad—. Tenemos que asumir que cualquier persona en la población puede ser escogida con igual probabilidad para formar parte de nuestra muestra: recordemos la analogía de Gallup sobre mover bien la sopa antes de probarla. Y hemos visto que si queremos hacer inferencias estadísticas sobre aspectos desconocidos o predicciones, entonces nuestras conclusiones siempre serán algo inciertas.

En el último capítulo vimos cómo podríamos usar el método de bootstrap para ver cuánta variación podemos esperar que tengan nuestros estadísticos de resumen si repitiésemos el proceso de muestreo una y otra vez, y después usásemos esta variabilidad para expresar nuestra incertidumbre sobre las características auténticas, pero desconocidas, de la población. De nuevo, esto solo requiere echar mano de la idea de «escoger aleatoriamente», una idea que incluso los niños pequeños entenderían como justa.



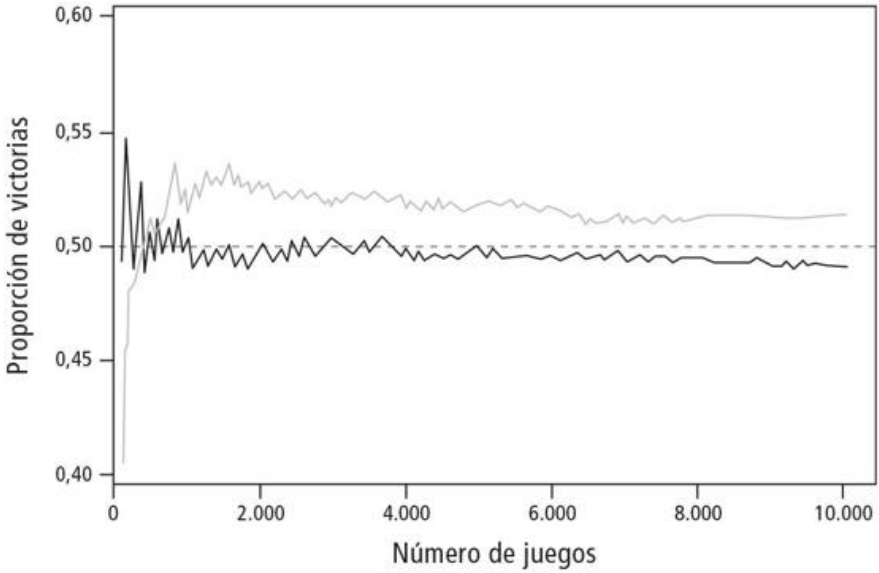


Figura 8.1. Simulación por ordenador de 10.000 repeticiones de los dos juegos. En el juego 1, ganas si sacas un seis en un máximo de 4 tiradas de dado; en el juego 2, ganas si sacas un seis doble en un máximo de 24 tiradas de dos dados. A lo largo de los primeros 100 juegos de cada uno de ellos (cuadro superior), parece que hay una mayor probabilidad de ganar en el juego 2, pero después de miles de rondas (cuadro inferior), queda claro que el juego 1 proporciona una apuesta marginalmente mejor.

Los cursos de Estadística empiezan tradicionalmente con la probabilidad —así lo hago yo cuando doy clases en Cambridge—, pero esta iniciación bastante matemática puede ser un obstáculo en la comprensión de todas las importantes ideas de los artículos precedentes, que no requerían la teoría de la probabilidad. Este libro, sin embargo, es parte de lo que podríamos denominar «nueva ola» de enseñanza de la estadística, en la que la teoría probabilística formal como base para la inferencia estadística no aparece hasta mucho más tarde.[111] Hemos visto que la simulación por ordenador es una herramienta muy poderosa tanto para explorar posibles eventos futuros como para hacer bootstrap con datos históricos, pero es una forma más bien rudimentaria y forzada de llevar a cabo un análisis estadístico. De manera que, aunque hemos avanzado mucho sin necesidad de recurrir a la teoría formal de la probabilidad, es ya hora de enfrentarnos a su papel vital con respecto al «lenguaje de la incertidumbre».

Pero ¿por qué esa reticencia a usar esa brillante teoría desarrollada a

lo largo de los últimos trescientos cincuenta años? A menudo me preguntan por qué la gente tiende a encontrar la probabilidad una idea difícil y no muy intuitiva, y yo contesto que, después de cuarenta años investigando y enseñando, he llegado a la conclusión de que se debe a que la probabilidad es una idea difícil y no muy intuitiva. Aquellos que encuentran que la probabilidad es compleja tienen toda mi simpatía. Incluso después de décadas de trabajo como experto en estadística, cuando se me hace una pregunta de colegio sobre probabilidad, tengo que sentarme en silencio con bolígrafo y papel, intentarlo de formas distintas y anunciar finalmente lo que espero que sea la respuesta correcta.

Empecemos con mi técnica de resolución de problemas favorita, que podría haberles ahorrado a los políticos más de una situación embarazosa.

Las reglas de la probabilidad presentadas, posiblemente, de manera un poco más sencilla

En 2012, se les hizo la siguiente pregunta a 97 miembros del Parlamento de Londres: «Si tiras una moneda dos veces, ¿cuál es la probabilidad de sacar dos caras?». La mayoría, 60 de 97, no supo dar la respuesta correcta.[112] ¿Cómo podrían estos políticos haberlo hecho mejor?

Quizá deberían haber sabido las reglas de la probabilidad, pero la mayoría de la gente no las sabe. Una alternativa habría sido usar una idea más intuitiva, que, en muchos experimentos de psicología, ha demostrado mejorar el razonamiento de la gente sobre la probabilidad.

Esta idea es la de la «frecuencia esperada». Cuando te enfrentas al problema de las dos monedas, te preguntas: «¿Qué esperaría que pasase si intentase ese experimento un cierto número de veces?». Digamos que intenta primero lanzar una moneda, y después otra, un total de cuatro veces. Sospecho que incluso un político podría, con solo pensar un poco, esperar obtener los resultados mostrados en la figura 8.2.

De manera que 1 de cada 4 veces esperaría usted obtener dos caras. Por tanto, siguiendo este razonamiento, la probabilidad de que en un intento concreto obtuviese dos caras es 1 de 4, o ¼. Que, afortunadamente, es la respuesta correcta.

Este árbol de frecuencia esperada puede ser transformado en un «árbol

de probabilidad», etiquetando cada «bifurcación» con la fracción de las ocasiones en las que la seguimos (figura 8.3). Debería quedar claro que la probabilidad global de una «rama» entera de un árbol de, digamos, una cara seguida de otra cara, se obtiene multiplicando las fracciones en las bifurcaciones a lo largo de la rama, tal que ½ × ½ = ¼.

Los árboles de probabilidad son una forma muy extendida y extremadamente efectiva de enseñar probabilidad en la escuela. De hecho, podemos usar este ejemplo sencillo de lanzar dos monedas para ver todas las reglas de la probabilidad, dado que el árbol de probabilidad muestra que:

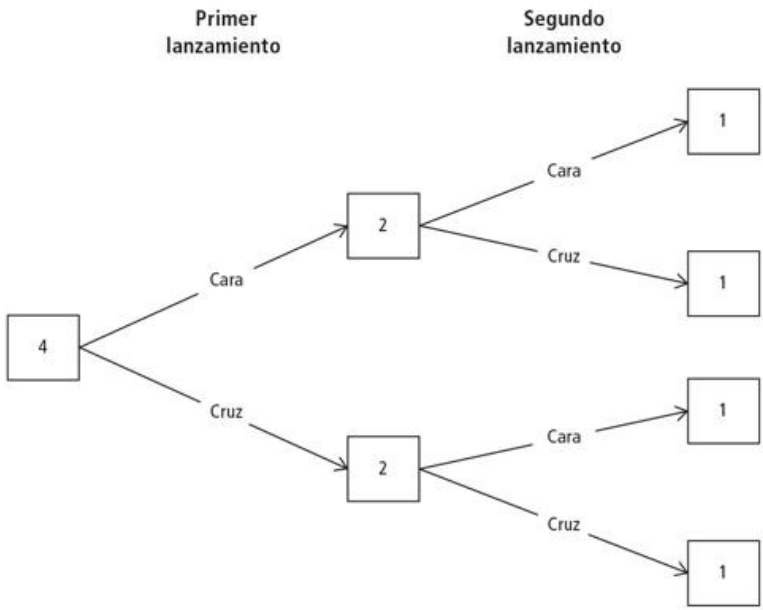


Figura 8.2. Árbol de frecuencia esperada para dos lanzamientos de monedas repetidos cuatro veces. Por ejemplo, se esperaría que dos de los cuatro primeros lanzamientos sean cara, y después que una de cada una de ellas sea cara y cruz en el segundo lanzamiento.

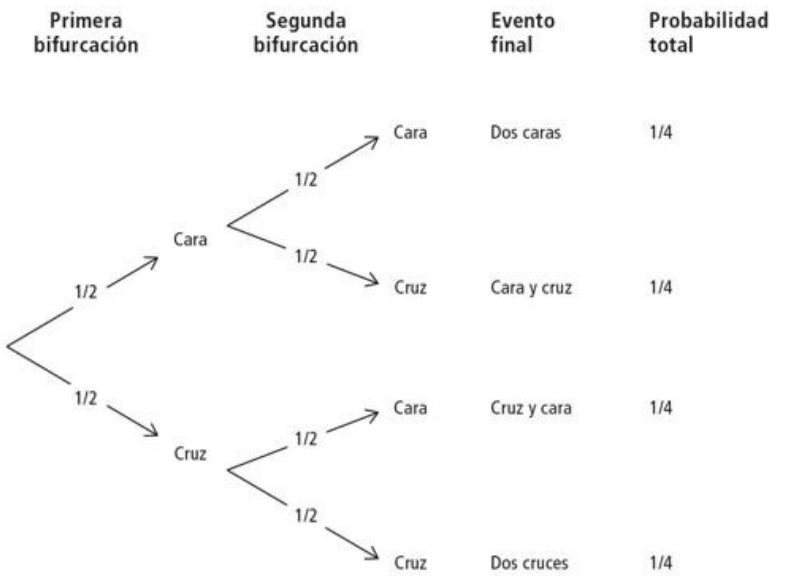


Figura 8.3. Árbol de probabilidad para lanzamientos de dos monedas. Cada «bifurcación» es etiquetada con la fracción de ocasiones en las que se toma. La probabilidad de una «rama» entera del árbol se obtiene multiplicando las fracciones de las bifurcaciones a lo largo de la rama.

1. La probabilidad de un evento es un número entre 0 y 1. Cero para eventos imposibles (por ejemplo, no sacar ni cara ni cruz), 1 para eventos seguros (sacar cualquiera de las cuatro combinaciones posibles).

2. Regla del complemento. La probabilidad de que ocurra un evento es 1 menos la probabilidad de que no ocurra. Por ejemplo, la probabilidad de «al menos una cruz» es 1 menos la probabilidad de «dos caras»: $1 - ¼ = ¾$.

3. La regla de la suma, o la regla de O. Sumamos las probabilidades de eventos mutuamente exclusivos (lo que significa que no pueden ocurrir ambos al mismo tiempo) para obtener la probabilidad total. Por ejemplo, la probabilidad de «al menos una cara» es ¾, dado que incluiría «dos caras» O «cara + cruz» O «cruz + cara», cada una de ellas con una probabilidad de ¼.

4. La regla de la multiplicación, o de Y. Multiplicamos las probabilidades para obtener la probabilidad global de la ocurrencia de una secuencia de eventos independientes (lo que significa que uno no afecta al otro). Por ejemplo, la probabilidad de una cara Y una cara es

½ × ½ = ¼.

Estas reglas básicas nos permiten solventar el problema de juego del Chevalier de Méré, revelando que tiene ciertamente una probabilidad del 52 % de ganar el juego 1, y un 49 % de probabilidades de ganar el juego 2.[113]

Aún estamos dando por hecho muchas cosas, incluso en este sencillo ejemplo de lanzar una moneda. Estamos asumiendo que la moneda es ecuánime y equilibrada, que se lanza apropiadamente (de manera que el resultado no sea predecible), que no cae de canto, que no cae un meteorito después del primer lanzamiento y así sucesivamente. Son consideraciones razonables (excepto posiblemente la del meteorito): sirven para enfatizar que todas las probabilidades que empleamos son condicionales —no existe la probabilidad incondicional de un evento —; siempre se asume que hay otros factores que podrían afectar a la probabilidad. Y, como veremos ahora, necesitamos ser cuidadosos acerca de a qué condicionamos la probabilidad.

Probabilidad condicional: cuando nuestras probabilidades dependen de otros eventos

Cuando se hace una revisión por cáncer de mama, la mamografía es aproximadamente un 90 % precisa, en el sentido de que el 90 % de las mujeres con cáncer, y el 90 % de las mujeres sin cáncer, serán correctamente clasificadas. Supongamos que el 1 % de las mujeres revisadas tengan realmente cáncer: ¿cuál es la probabilidad de que una mujer escogida aleatoriamente tenga una mamografía positiva, y si la tiene, cuál es la probabilidad de que realmente tenga cáncer?

En el caso de las dos monedas, los eventos eran independientes, en el sentido de que la probabilidad de sacar cara en el segundo lanzamiento no dependía de lo que se haya sacado en el primero. En el colegio normalmente aprendemos acerca de eventos dependientes con respuestas a preguntas algo tediosas sobre, digamos, una serie de calcetines de diferentes colores extraídos de un armario. El ejemplo mencionado arriba es algo más relevante para la vida real.

Este tipo de problema es un clásico en los test de inteligencia y no es fácil de resolver, pero usando la idea de frecuencias esperadas, se vuelve sorprendentemente sencillo. Lo crucial es pensar qué esperaríamos que le ocurra a un gran grupo de mujeres, digamos 1.000, como se muestra en la figura 8.4.

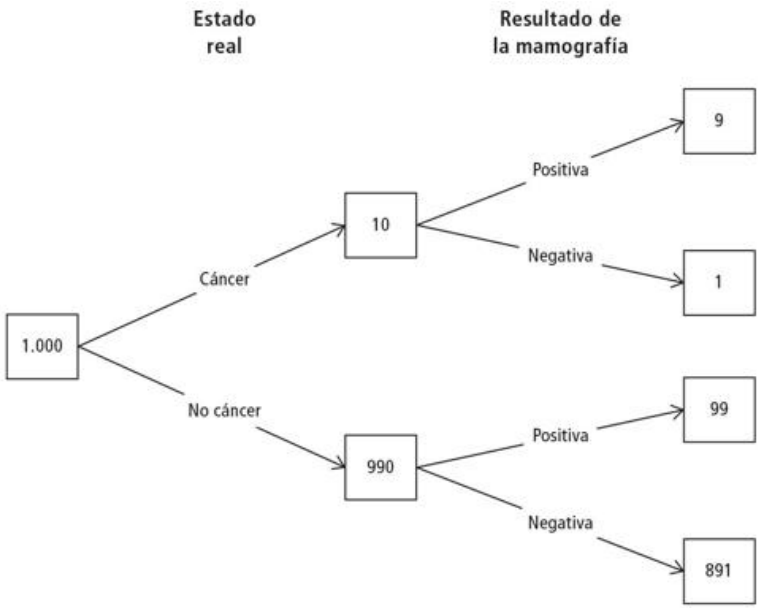


Figura 8.4. Árbol de frecuencia esperada que muestra lo que esperaríamos que les ocurriese a 1.000 mujeres que son revisadas por cáncer de mama. Asumimos que un 1 % de las mujeres tienen cáncer de mama, y que la mamografía clasifica correctamente al 90 % de las mujeres con cáncer, y al 90 % de las mujeres sin cáncer. En general, esperaríamos 9 + 99 = 108 mamografías positivas, de las cuales 9 realmente tendrían cáncer.

De las 1.000 mujeres, 10 (1 %) tienen realmente cáncer de mama. De estas 10, 9 (90 %) tienen un resultado positivo. Pero de 990 mujeres sin cáncer, 99 (10 %) obtienen un resultado positivo falso de su mamografía. Juntando todo ello, hay un total de 9 + 99 = 108 mamografías positivas, y, así, la probabilidad de que una mujer escogida aleatoriamente obtenga un resultado positivo es de 108/1.000, o alrededor del 11 %. Pero de estas 108, solo 9 tienen realmente cáncer, de manera que solo hay una probabilidad de 9/108 = 8 % de que la mujer tenga realmente cáncer.

Este ejercicio en probabilidades condicionales nos ayuda a comprender un resultado muy contraintuitivo: a pesar de la «precisión del 90 %» del escáner, una gran mayoría de mujeres con una mamografía positiva no tiene cáncer de mama. Es fácil confundir la probabilidad de un test positivo, dado que se tiene cáncer, con la probabilidad de cáncer, dado un test positivo.

Este tipo de confusión es popularmente conocida como falacia del

fiscal, por su prevalencia en casos judiciales que implican ADN. Un experto forense podría afirmar, por ejemplo, que «si el acusado es inocente, hay solo una probabilidad entre mil millones de que coincida con el ADN encontrado en la escena del crimen». Pero esto se interpreta equivocadamente como que «dada la evidencia del ADN, solo hay una probabilidad de entre mil millones de que el acusado sea inocente».[114]

Es este un error fácil de cometer, pero la lógica es tan defectuosa como pasar de la afirmación según la cual «si eres el papa, entonces eres católico» a «si eres católico, entonces eres el papa»; en este caso, el error es más fácil de identificar.

Pero ¿qué es «probabilidad»?

En la escuela nos enseñan las matemáticas de la distancia, el peso y el tiempo, que podemos medir con una regla, una báscula o un reloj. Pero ¿cómo medimos la probabilidad? No existe el «probabilómetro». Es como si la probabilidad fuese alguna cantidad «virtual», a la que le podemos asignar un número, pero que nunca podemos medir directamente.

Aún más preocupante es la siguiente cuestión, bastante obvia: ¿qué significa la probabilidad? ¿Cuál sería una buena definición? Esto podría resultar algo pedante, pero la filosofía de la probabilidad es un tema apasionante en sí mismo, y al mismo tiempo juega un gran papel en las aplicaciones prácticas de la estadística.

No esperemos un consenso puro entre los «expertos». Pueden estar de acuerdo sobre las matemáticas de la probabilidad, pero los filósofos y los estadísticos han desarrollado todo tipo de ideas diferentes sobre qué significan estos elusivos números, y han discutido intensamente sobre ellos. Algunas sugerencias populares son las siguientes:

→ Probabilidad clásica. La que nos enseñaron en la escuela, basada en la simetría de monedas, dados, tacos de cartas y demás, y que podría definirse como «el ratio del número de resultados que favorecen un evento dividido por el número total de resultados posibles, asumiendo que los resultados son igualmente probables». Por ejemplo, la probabilidad de sacar un «uno» con un dado equilibrado es de 1/6, dado que tiene seis caras. Pero esta definición es un poco circular, dado que necesitamos tener una definición de qué es «igualmente probable».

→ Probabilidad «enumerativa».[115] Supongamos que hay tres

calcetines blancos y cuatro calcetines negros en un cajón, y cogemos un calcetín de manera aleatoria. ¿Cuál es la probabilidad de coger un calcetín blanco? Es 3/7, que se obtiene enumerando las probabilidades. Muchos de nosotros tuvimos que sufrir preguntas como esta en la escuela; se trata esencialmente de una extensión de la idea clásica discutida más arriba, que requiere la noción de una «elección aleatoria» de un conjunto físico de objetos. Ya hemos usado esta idea con frecuencia, cuando describimos un punto de datos extraído aleatoriamente de una población.

→ Probabilidad de «frecuencia a largo plazo». Se basa en el porcentaje de veces en que ocurre un evento en una secuencia infinita de experimentos idénticos, exactamente lo que encontramos cuando simulamos los juegos del Chevalier. Esto podría ser razonable (al menos teóricamente) para eventos repetidos infinitamente, pero ¿qué ocurre en el caso de eventos únicos como una carrera de caballos, o el tiempo que hará mañana? De hecho, casi cualquier situación realista no es, ni siquiera en principio, infinitamente repetible.

→ Propensión o «destino». Esta es la idea de que una situación tiene alguna tendencia objetiva a producir un evento. Esto es en principio atractivo —si fuese usted un ser omnisciente, quizá podría decir que hay una probabilidad concreta de que su autobús llegue pronto o de ser atropellado hoy por un coche—. Pero no parece proporcionar un fundamento para que nosotros, simples mortales, estimemos el un tanto metafísico «destino verdadero».

→ Probabilidad subjetiva o «personal». Se trata de un juicio específico de una persona sobre una ocasión específica, basado en su conocimiento actual, y se interpreta aproximadamente en términos de las probabilidades de ganar (cuando hay poco en juego) que esa persona encontraría razonables. Por ejemplo, si me dan 1 libra si puedo hacer malabares con tres bolas durante cinco minutos, y estoy dispuesto a ofrecer 60 peniques (no reembolsables) por la apuesta, entonces mi probabilidad para ese evento es 0,6.

Los distintos «expertos» tienen sus propias preferencias acerca de estas alternativas, pero personalmente yo prefiero la última interpretación —probabilidad subjetiva—. Esto significa que en mi opinión cualquier probabilidad numérica está esencialmente construida de acuerdo con lo que sé en cada situación concreta —de hecho, la probabilidad no «existe» realmente (excepto posiblemente a nivel subatómico)—. Este enfoque forma la base de la escuela bayesiana de inferencia estadística, que exploraremos en detalle en el capítulo 11.

Pero, afortunadamente, no tiene que estar de acuerdo con mi (bastante controvertida) posición según la cual las probabilidades numéricas no existen objetivamente. Basta con asumir que las monedas y otros dispositivos de aleatorización son objetivamente aleatorios, en el sentido de que generan datos que son tan impredecibles como para ser indistinguibles de los que esperaríamos de probabilidades «objetivas». De manera que, en general, actuamos como si las observaciones fueran aleatorias, incluso aunque sabemos que esto no es estrictamente así. Los ejemplos más extremos de ello son los generadores de números falsamente aleatorios, que de hecho se basan en cálculos lógicos y completamente predecibles. No contienen ningún tipo de aleatoriedad, pero su mecanismo es tan complejo que, en la práctica, es indistinguible de las secuencias auténticamente aleatorias (por ejemplo, aquellas que se obtienen a partir de una fuente de partículas subatómicas).[116]

Esta capacidad algo extraña para actuar como si algo fuese verdad, cuando se sabe que realmente no lo es, se consideraría normalmente como algo peligrosamente irracional. No obstante, resulta de lo más útil cuando se trata de usar la probabilidad como la base para el análisis estadístico de los datos.

Llegamos ahora al estadio crucial (pero difícil de establecer) de la conexión general entre la teoría de la probabilidad, los datos y el aprendizaje acerca de la población objetivo en la que estamos interesados.

La teoría de la probabilidad juega un papel en lo que denominaremos situación 1:

1. Cuando puede considerarse que el punto de datos ha sido generado por algún mecanismo de aleatorización, por ejemplo lanzar un dado, una moneda o asignar aleatoriamente a un individuo a un tratamiento médico usando un generador de números falsamente aleatorios, y registrando después los resultados de los tratamientos.

Pero en la práctica, nos enfrentamos a la situación 2:

2. Cuando un punto de datos preexistente es escogido mediante un mecanismo de aleatorización, digamos, por ejemplo, cuando seleccionamos personas para formar parte de una encuesta.

Y en muchas ocasiones, nuestros datos se derivan de la situación 3:

3. Cuando no hay nada aleatorio, pero actuamos como si el punto de datos hubiese sido realmente generado por algún proceso aleatorio,

por ejemplo, al interpretar el peso al nacer del bebé de nuestra amiga.

La mayor parte de las exposiciones no aclaran estas distinciones: la probabilidad se enseña generalmente usando mecanismos de aleatorización (situación 1) y la estadística se enseña a través de la idea del «muestreo aleatorio» (situación 2), pero, de hecho, la mayoría de las aplicaciones de la estadística no implican ningún mecanismo o muestreo aleatorios (situación 3).

Pero consideremos primero las situaciones 1 y 2. Justo antes de poner en marcha el mecanismo de aleatorización, asumimos que tenemos un conjunto de resultados posibles que podrían ser observados, junto con sus probabilidades correspondientes —por ejemplo, una moneda puede salir cara o cruz, teniendo cada resultado una probabilidad de ½—. Si asociamos cada uno de estos posibles resultados con una cantidad, por ejemplo, en este caso, 0 para cruz y 1 para cara, entonces diremos que tenemos una variable aleatoria con una distribución de probabilidad. En la situación 1, el mecanismo de aleatorización asegura que la observación es generada de manera aleatoria a partir de la distribución, y cuando es observada, la aleatoriedad desaparece y todos esos futuros potenciales colapsan en la observación real.[117] Igualmente, en la situación 2, si seleccionamos un individuo de manera aleatoria y, por ejemplo, medimos sus ingresos, entonces esencialmente lo que haremos es extraer una observación de manera aleatoria de una distribución poblacional de ingresos.

De manera que la probabilidad es claramente relevante cuanto tenemos un mecanismo de aleatorización. Pero la mayor parte del tiempo simplemente consideramos todas las medidas que tenemos disponibles en ese momento, que pueden haber sido reunidas informalmente o, como vimos en el capítulo 3, representar toda observación posible: pensemos en las tasas de supervivencia para la cirugía infantil de corazón en diferentes hospitales o en todos los resultados de los exámenes realizados por niños británicos: en ambos casos, tenemos todos los datos disponibles, y no ha habido muestreo aleatorio.

En el capítulo 3 discutimos la idea de una población metafórica, que incluiría todas las eventualidades posibles que podrían haber ocurrido, pero no lo han hecho. Ahora necesitamos prepararnos para un paso aparentemente irracional: necesitamos actuar como si los datos fueran generados por un mecanismo aleatorio a partir de esa población, aunque sepamos que no ha sido así.

Si observamos todo, ¿qué espacio queda para la probabilidad?

¿Con cuánta frecuencia esperamos que se produzcan siete o más homicidios en Inglaterra y Gales en un solo día?

Cuando se suceden eventos extremos muy próximos unos de otros, como por ejemplo múltiples accidentes de avión o desastres naturales, hay una propensión natural a pensar que están relacionados en alguna medida. En estos casos es importante establecer hasta qué punto esos eventos son inusuales, y el siguiente ejemplo muestra cómo hacerlo.

Para evaluar lo rara que resulta una serie de al menos siete homicidios en un día, podemos examinar datos de tres años (1.095 días) entre abril de 2013 y marzo de 2016, en los que se produjeron 1.545 incidentes homicidas en Inglaterra y Gales, una media de 1.545/1.095 = 1,41 al día.[118] A lo largo de este período no hubo ningún día con siete o más incidentes, pero sería muy ingenuo concluir a partir de ahí que es imposible que ocurra algo así. Si podemos desarrollar una distribución de probabilidad razonable para el número de homicidios al día, entonces podemos responder a la pregunta que nos planteábamos.

Pero ¿cuál es la justificación para desarrollar una distribución de probabilidad? El número de homicidios registrados cada día en un país es simplemente un hecho —no ha habido muestreo y no hay un elemento aleatorio explícito que genere cada evento desafortunado—. Solo un mundo inmensamente complejo e impredecible. Pero sea cual sea nuestra filosofía personal acerca de la suerte, lo cierto es que es útil actuar como si estos eventos fueran producidos a través de algún proceso aleatorio determinado por la probabilidad.

Resultaría útil imaginar que al comienzo de cada día tenemos una gran población de personas, cada una de las cuales tiene una posibilidad muy pequeña de ser víctima de un homicidio. Datos de este tipo pueden ser representados como observaciones de una distribución Poisson, que fue desarrollada originalmente por Siméon Denis Poisson en Francia en la década de 1830 para representar la pauta de condenas erróneas por año. Desde entonces se ha empleado para modelizar cualquier cosa, desde el número de goles marcados por un equipo de fútbol en un partido hasta el número de oficiales prusianos muertos anualmente por una coz propinada por sus caballos. En cada una de estas situaciones hay un gran número de oportunidades para que se produzca un evento, pero cada una de ellas tiene una probabilidad de ocurrencia muy baja, y esto da lugar a la extraordinariamente versátil distribución Poisson.

Mientras que la distribución normal (o de Gauss) del capítulo 3 requería dos parámetros —la media y la desviación típica de la población—, la distribución Poisson depende solo de la media. En el caso de nuestro ejemplo actual se trata del número esperado de incidentes homicidas cada día, que asumimos que es 1,41, el número diario medio a lo largo de este período de tres años. Deberíamos, por tanto, comprobar cuidadosamente si la Poisson es una asunción razonable, de manera que sea razonable actuar como si el número de homicidios diarios fuese una observación aleatoria extraída de una distribución Poisson con media 1,41.

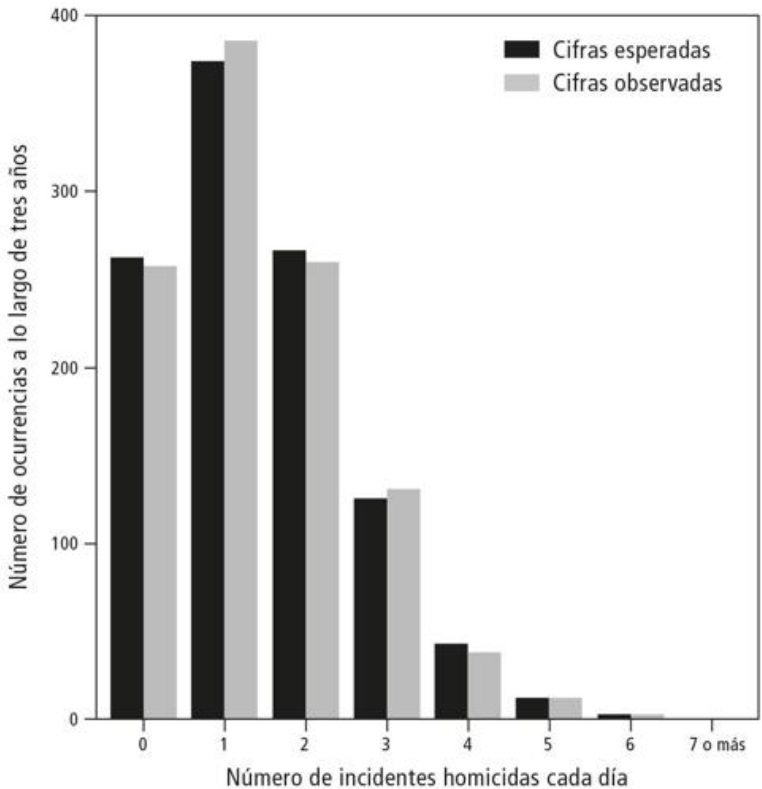


Figura 8.5. Número diario observado y esperado (asumiendo una distribución Poisson) de incidentes homicidas registrados entre 2013 y 2016 en Inglaterra y Gales. (Comparaciones de recuentos de homicidios al día con una distribución Poisson: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/compendium/focusonviolentcrimeandsexualoffences/yearendingmarch2016/homicide#statistical-interpretation-of-trends-in-homicides).

Por ejemplo, solo sabiendo esa media, podemos usar la fórmula de la

distribución Poisson, o bien un software estándar, para calcular que habría una probabilidad de 0,01134 de que se produzcan exactamente 5 homicidios en un día, lo que significa que a lo largo de 1.095 días esperaríamos $1.095 \times 0{,}01134 = 12{,}4$ días en los cuales hubiese exactamente 5 incidentes homicidas. Sorprendentemente, el número de días a lo largo de un período de tres años en los que hubo cinco homicidios fue de… trece.

La figura 8.5 compara la distribución esperada del número diario de incidentes homicidas basada en una función Poisson, y la distribución empírica de datos real a lo largo de estos 1.095 días —las cifras son ciertamente muy parecidas—. En el capítulo 10 veremos cómo comprobar formalmente si la asunción Poisson está justificada.

En respuesta a la pregunta planteada al comienzo de esta sección, podemos calcular a partir de la distribución Poisson la probabilidad de obtener siete o más incidentes en un día, que resulta ser de 0,07 %, lo que significa que podemos esperar que un evento así ocurra de media cada 1.535 días, o aproximadamente una vez cada cuatro años. Podemos concluir que este evento es bastante poco probable, pero no imposible.

El ajuste de esta distribución matemática de probabilidades a los datos empíricos es casi preocupantemente bueno. Y ello aun teniendo en cuenta que hay una historia única detrás de cada uno de estos trágicos sucesos, la mayoría de los cuales es impredecible, dado que los datos se comportan como si realmente hubiesen sido generados por algún mecanismo aleatorio conocido. Una forma de verlo es pensar que otras personas podrían haber sido asesinadas, pero no lo fueron; hemos observado uno de los mundos posibles que podrían haber ocurrido, igual que cuando lanzamos una moneda observamos una de las muchas posibles secuencias.

Adolphe Quetelet era un astrónomo, estadístico y sociólogo belga de mediados del siglo XIX, y fue uno de los primeros que llamaron la atención sobre la sorprendente predictibilidad de las pautas globales derivadas de eventos individualmente impredecibles. Estaba intrigado por la presencia de distribuciones normales en fenómenos naturales, como la distribución del peso al nacer que vimos en el capítulo 3, y acuñó la idea de «l'homme moyen» (el hombre medio), que tenía el valor medio en todas estas características. Desarrolló la idea de la «física social», basada en el hecho de que la regularidad de las estadísticas sociales parecía reflejar un proceso subyacente casi mecanicista. Así como las moléculas aleatorias de un gas se juntan para generar propiedades físicas predecibles, las acciones

impredecibles de millones de vidas individuales se juntan para producir, por ejemplo, tasas nacionales de suicidio que apenas cambian de año en año.

Afortunadamente no tenemos que creer que los eventos están determinados por la pura aleatoriedad (sea esta la que sea). Se trata simplemente de asumir que el «azar» juega un papel que engloba toda la inevitable impredictibilidad del mundo, lo que en ocasiones se denomina variabilidad natural. Hemos establecido, por tanto, que la probabilidad constituye el fundamento matemático apropiado tanto para la aleatoriedad «pura», que se da en las partículas subatómicas, las monedas, los dados, etc.; y la variabilidad «natural», inevitable, como la referida al peso al nacer, la supervivencia después de la cirugía, los resultados de los exámenes, los homicidios y cualquier otro fenómeno que no sea totalmente predecible.

En el capítulo siguiente nos ocuparemos de un avance fundamental en la historia del conocimiento humano: cómo pueden combinarse estos dos aspectos de la probabilidad para proporcionar una base rigurosa a la inferencia estadística formal.

Resumen

→ La teoría de la probabilidad proporciona un lenguaje formal y matemático para enfrentarnos a los fenómenos aleatorios.

→ Las implicaciones de la probabilidad no son intuitivas, pero pueden mejorarse usando la idea de las frecuencias esperadas.

→ La idea de la probabilidad es útil incluso cuando no hay un uso explícito de un mecanismo de aleatorización.

→ Muchos fenómenos sociales muestran una regularidad notable en su pauta general, mientras que los eventos individuales son enteramente impredecibles.

[110] Para el juego 1, pensó que cuatro lanzamientos, con una probabilidad de 1/6 cada vez, tendrían una probabilidad global de ganar de 4 veces 1/6 o, lo que es lo mismo, de 2/3. Igualmente, para el juego 2, pensó que 24 lanzamientos con una probabilidad de 1/36 cada vez significaba un 24/36 = probabilidad de 2/3, la misma que el juego 1. Se trata de un error que aún cometen habitualmente los estudiantes; para mostrar que este cálculo no puede estar bien, considérese simplemente lo siguiente: si hubiese lanzado 12 veces en

el juego 1, ¿sería la probabilidad de ganar 12 veces 1/6 = 2? El razonamiento correcto se provee en la nota 4 de este capítulo.

[111] Para una discusión al respecto, así como para encontrar herramientas de métodos de simulación para enseñar estadística, véase M. Pfannkuch et al., «Bootstrapping Students' Understanding of Statistical Inference», TLRI (2013), y K Lock Morgan et al., «STATKEY: Online Tools for Bootstrap Intervals and Randomization Tests», ICOTS 9 (2014).

[112] Alerta de spoiler: la respuesta es ¼, o 25 %, o 0,25.

[113] El razonamiento correcto es el siguiente. Consideremos el juego 1. Hay muchas formas de ganar, pero solo una de perder: sacando cuatro números seguidos distintos de seis. Es por tanto más sencillo encontrar la probabilidad de perder (este truco es común). La probabilidad de sacar un número distinto de 6 es 1 − 5/6 = 5/6 (regla del complemento), y la probabilidad de sacar cuatro números seguidos distintos de seis es 5/6 × 5/6 × 5/6 = (5/6)4 = 625/1296 = 0,48 (regla de la multiplicación). De manera que la probabilidad de ganar es de 1 − 0,48 = 0,52 (otra vez la regla del complemento). Un razonamiento similar para el juego 2 lleva a una probabilidad de ganar de 1 − (35/36)24 = 0,49, lo que muestra que el juego 1 es ligeramente más favorable. Estas normas muestran también el error en el razonamiento del Chevalier: estaba sumando las probabilidades de eventos que no eran mutuamente exclusivos. A través de ese razonamiento la probabilidad de un seis tras lanzar un dado 12 veces sería de 12 / 6 = 2, lo que no tiene mucho sentido.

[114] Esto también se conoce como «ley de la transitividad del condicional», que suena deliciosamente oscura, pero que significa simplemente que la probabilidad de A dado B se confunde con la probabilidad de B dado A.

[115] Agradezco a Philip Dawid por, aparentemente, haber inventado este término.

[116] Esto parte del supuesto de que el generador de números falsamente aleatorios está bien diseñado, y que el uso previsto del mismo es la modelización estadística o algo similar. No son suficientemente buenos para aplicaciones criptográficas, donde se puede usar la predictibilidad para descifrar un encriptado.

[117] Esto se puede considerar análogo a la mecánica cuántica, en la que el estado actual de, por ejemplo, un electrón, se define como una

función de onda que colapsa a un estado único cuando es realmente observada.

[118] Un «incidente homicida» es cuando se sospecha que la misma persona (o grupo de personas) ha cometido uno o más homicidios relacionados. De manera que los tiroteos en masa o atentados terroristas contarían como un solo incidente.

09

Juntar la probabilidad

y la estadística

Advertencia. Este es quizás el capítulo más exigente del libro, pero perseverar en este importante tema le dará una

valiosa comprensión de la inferencia estadística.

En una muestra aleatoria de 100 personas, encontramos que 20 son zurdas. ¿Qué podemos decir sobre el porcentaje de la población que es zurdo?

En el capítulo anterior discutimos la idea de variable aleatoria: un único punto de datos extraído de una distribución de probabilidad descrita por parámetros. Pero no estamos interesados en un único punto de datos —normalmente tenemos una masa de datos que resumimos recurriendo a la media, la mediana y otros estadísticos—. El paso fundamental que daremos en este capítulo es considerar que esos mismos estadísticos son variables aleatorias, extraídas de sus propias distribuciones.

Es este un gran avance, que no solo ha supuesto un desafío para generaciones de estudiantes de Estadística, sino también para generaciones de expertos en estadística que han intentado determinar de qué distribuciones hemos de asumir que se derivan estos estadísticos. Y, teniendo en cuenta la discusión sobre el bootstrap en el capítulo 7, sería razonable preguntar por qué necesitamos todas estas matemáticas, cuando podemos determinar intervalos de incertidumbre y todo lo demás utilizando el enfoque de simulación basado en el bootstrap. Por ejemplo, la cuestión planteada al comienzo de este capítulo podría responderse tomando nuestros datos observados de 20 individuos zurdos y 80 diestros, haciendo un re-muestreo repetido de 100 observaciones de esta base de datos, con reemplazamiento, y observando la distribución del porcentaje observado de personas zurdas.

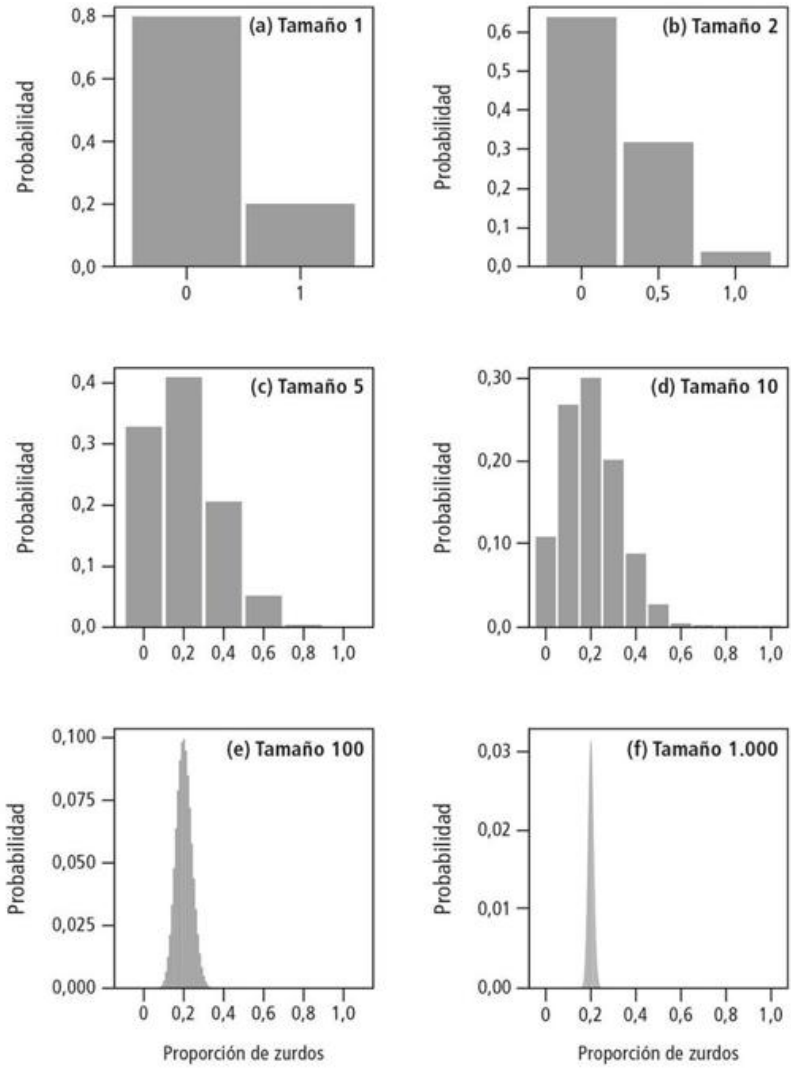


Figura 9.1. Distribución de probabilidad del porcentaje observado de zurdos en muestras aleatorias de 1, 2, 5, 10, 100 y 1.000 personas, en las que el verdadero porcentaje de zurdos en la población es de 0,2. La probabilidad de obtener al menos un 30 % de zurdos en la muestra se obtiene sumando todas las probabilidades de las barras a la derecha de 0,3.

Pero estas simulaciones son rudimentarias y llevan mucho tiempo, especialmente con grandes bases de datos, y en circunstancias más complejas no está muy claro qué es lo que deberíamos simular. Frente a ello, las fórmulas derivadas de la teoría de la probabilidad proporcionan al mismo tiempo información y conveniencia, y siempre

generan la misma respuesta, dado que no dependen de una simulación particular. La parte mala es que esta teoría se basa en asunciones, y no deberíamos dejarnos llevar por la impresionante álgebra que la acompaña y aceptar conclusiones que no estén justificadas. Nos ocuparemos de ello con detalle más adelante, pero, primero, habiendo ya apreciado lo valiosas que resultan la normal y la Poisson, necesitamos introducir otra importante distribución de probabilidad.

Supongamos que extraemos muestras de diferentes tamaños de una población que contiene exactamente un 20 % de zurdos y un 80 % de diestros, y calculamos la probabilidad de observar porcentajes posibles de zurdos. Por supuesto, esta es la forma errónea de enfrentarse al problema —queremos usar la muestra conocida para aprender acerca de la población desconocida— pero solo podemos llegar a esta conclusión explorando primero cómo una población conocida da lugar a diferentes muestras.

El caso más sencillo es una muestra de uno, en cuyo caso el porcentaje observado debe ser 0 o 1, dependiendo de si seleccionamos a un diestro o a un zurdo; y estos eventos se producen con una probabilidad de 0,8 y 0,2 respectivamente. La distribución de probabilidad resultante se muestra en la figura 9.1(a).

Si seleccionamos dos individuos aleatoriamente, entonces el porcentaje de zurdos será o bien 0 (ambos diestros), 0,5 (uno de cada) o 1 (ambos zurdos). Estos eventos ocurrirán con probabilidades 0,64, 0,32 y 0,04, respectivamente.[119] Esta distribución de probabilidad se expone en la figura 9.1(b). Igualmente, podemos usar la teoría de la probabilidad para calcular la distribución de probabilidad para los porcentajes observados de zurdos en las muestras de 5, 10, 100 y 1.000 personas, que se muestran en la figura 9.1. Estas distribuciones se basan en lo que se conoce como distribución binomial, y también pueden proporcionarnos la probabilidad de, por ejemplo, obtener al menos un 30 % de zurdos para una muestra de 100, lo que se conoce como la cola.

La media de un variable aleatoria se conoce también como su esperanza, y en todas estas muestras esperamos un porcentaje de 0,2, o un 20 %: todas las distribuciones que se muestran en la figura 9.1 tienen una media de 0,2. La desviación típica para cada una de ellas se obtiene por una fórmula que depende del porcentaje subyacente, en este caso 0,2, y del tamaño de la muestra. Nótese que la desviación típica de un estadístico se denomina normalmente error típico, para distinguirlo de la desviación típica de la distribución poblacional de la que se deriva.

La figura 9.1 tiene algunas características distintivas. En primer lugar, las distribuciones de probabilidad tienden a adoptar formas regulares, simétricas, normales, a medida que se incrementa el tamaño de la muestra, justo como observamos cuando usamos las simulaciones empleando bootstrap. En segundo lugar, las distribuciones se vuelven más estrechas a medida que el tamaño aumenta. El ejemplo siguiente muestra cómo se puede emplear una aplicación simple de estas ideas para identificar rápidamente si una afirmación estadística es razonable o no.

¿Es verdad que algunas áreas del Reino Unido tienen unas tasas de fallecimientos por cáncer colorrectal tres veces mayores que otras?

El titular de la respetada página web de noticias de la BBC en septiembre de 2011 era alarmante: «Una variación de 3 a 1 en las tasas de fallecimientos por cáncer colorrectal». El artículo explicaba que distintas áreas del Reino Unido tenían sorprendentes diferencias en tasas de fallecimientos por cáncer colorrectal, y un experto sugería que era «extremadamente importante que las organizaciones locales del NHS examinasen la información proveniente de sus propias áreas y la usasen para recomendar potenciales cambios en los servicios prestados».

Una diferencia de 3 a 1 suena extraordinariamente alta. Cuando el bloguero Paul Barden leyó el artículo, se preguntó: «¿Es cierto que personas que viven en distintas partes del país se enfrentan a diferencias tan grandes e importantes en su riesgo de morir de cáncer colorrectal? ¿Qué podía generar una discrepancia tan abultada?». Le pareció todo tan implausible que decidió investigar. Los datos estaban disponibles en Internet para todo el mundo, algo que ciertamente es digno de admiración, y encontró que, efectivamente, corroboraban lo afirmado por la BBC: en 2008 se produjo una variación de más de 3 a 1 entre las tasas de fallecimiento anuales de personas con cáncer colorrectal. Variaron desde las 9 personas por 100.000 en Rossendale, Lancashire, a 31 por 100.000 habitantes en la ciudad de Glasgow. [120]

Pero su investigación no acabó ahí. A continuación hizo un diagrama cruzando las tasas de fallecimientos con la población de cada distrito, lo que generó la forma que se muestra en la figura 9.2. Está claro que los puntos (excepto el ejemplo extremo de Glasgow) adoptan una cierta forma de embudo, en la que las diferencias entre los distritos se agrandan a medida que su población se hace más pequeña. Paul añadió entonces límites de control, que muestran dónde podríamos esperar que estuviesen los puntos si las diferencias entre las tasas

observadas se debiesen a la variabilidad natural e inevitable en el número de personas que mueren de cáncer colorrectal cada año, más que debido a una variación sistemática en el riesgo subyacente experimentado en los distintos distritos. Estos límites de control se obtienen basándose en la asunción de que el número de fallecidos por cáncer colorrectal en cada área son una observación de una distribución binomial con un tamaño muestral igual a la población adulta del área, y una probabilidad subyacente de 0,000176 de muerte por cáncer colorrectal al año: este es el riesgo individual medio para todo el país. Los límites de control se establecen para que incluyan el 95 % y el 99,8 % de la distribución de probabilidad, respectivamente. Este tipo de gráfico se denomina gráfico de embudo y se usa ampliamente cuando hay múltiples autoridades sanitarias o instituciones, ya que permite la identificación de valores atípicos sin crear clasificaciones espurias.

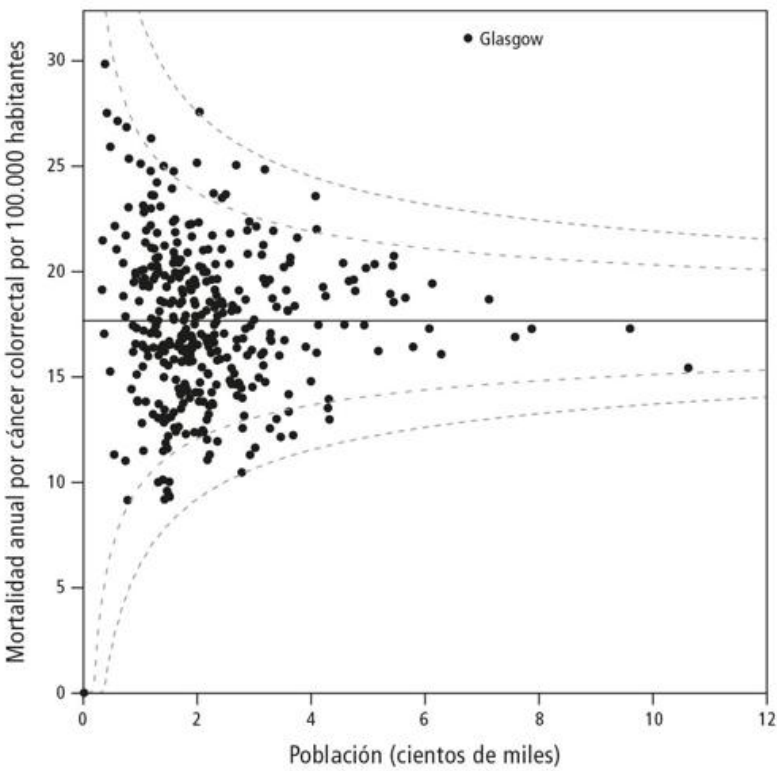


Figura 9.2. Tasas anuales de fallecimientos por cáncer colorrectal por 100.000 habitantes en 380 distritos del Reino Unido (excepto Gales), cruzados con la población del distrito. Los dos conjuntos de líneas discontinuas indican las zonas en las que esperaríamos que se encontrasen el 95 % y el 99,8 % de los distritos, si no hubiese

diferencias reales entre los riesgos, y se derivan de un supuesto de una distribución binomial subyacente. Solo Glasgow muestra cierta evidencia de un riesgo subyacente diferente de la media. Esta forma de observar los datos se denomina «gráfico de embudo».

Los datos se ajustan bastante bien a los límites de control, lo que significa que las diferencias entre distritos son esencialmente lo que esperaríamos si solo se tratase de una variabilidad aleatoria. Los distritos más pequeños tienen menos casos y, por ello, son más vulnerables a la pura aleatoriedad, y por ello tienden a tener resultados más extremos —la tasa de Rossendale se basaba en solo 7 fallecimientos, por lo que podía ser alterada radicalmente con solo unos pocos casos adicionales—. De manera que, a pesar del dramático titular de la BBC, no hay una gran exclusiva en este caso: esperaríamos una variabilidad de tres a uno en las tasas observadas, incluso aunque el riesgo subyacente en los distintos distritos fuese precisamente el mismo.

Hay una lección crucial que extraer de este sencillo ejemplo. Incluso en la era de los datos abiertos, la ciencia de los datos y el periodismo basado en datos, seguimos necesitando unos principios estadísticos básicos para no dejarnos engañar por unas pautas aparentes en los números.

Este gráfico revela que la única observación de interés es el dato atípico de la ciudad de Glasgow. ¿Es el cáncer colorrectal un fenómeno particularmente escocés? ¿Es este punto de datos realmente correcto? Datos más recientes para el período 2009-2011 revelan que la mortalidad por cáncer colorrectal en el Gran Glasgow era de 20,5 por 100.000 personas, en Escocia, de 19,6, y en Inglaterra, de 16,4. Estos datos plantean algunas dudas sobre el valor específico de la ciudad de Glasgow y muestran que Escocia tiene tasas mayores que Inglaterra. Como es lo habitual, las conclusiones derivadas de un ciclo de resolución de problemas plantean nuevas cuestiones, y, como consecuencia de ello, el ciclo vuelve a empezar.

El teorema central del límite

Los puntos de datos individuales pueden ser extraídos de una amplia variedad de distribuciones poblacionales, algunas de las cuales podrían estar muy sesgadas, con largas colas, como las referidas a los ingresos o las parejas sexuales. Pero ahora hemos realizado un giro crucial. Estamos utilizando distribuciones de estadísticos en lugar de puntos de datos individuales, y estos estadísticos son normalmente algún tipo de media. Ya hemos visto en el capítulo 7 que la

distribución de la media muestral de re-muestreos bootstrap tiende a tener una encomiable forma simétrica, sea cual sea la forma de la distribución original de los datos, y ahora podemos ir más allá, siguiendo una idea más profunda y bastante notable, establecida hace unos trescientos años.

El ejemplo de los zurdos muestra que la variabilidad en los porcentajes observados se hace más pequeña a medida que el tamaño de la muestra se incrementa —por eso el embudo de la figura 9.2 se estrecha alrededor de la media—. Esta es la clásica ley de los números grandes, desarrollada por el matemático suizo Jacob Bernoulli a comienzos del siglo XVIII; se denomina ensayo de Bernoulli a un único lanzamiento de una moneda, con valor 1 si sale cara y 0 si sale cruz, y adoptando una distribución de Bernoulli. Si se sigue lanzando una moneda equilibrada, realizando más y más ensayos de Bernoulli, el porcentaje de cada resultado se acercará cada vez más a un 50 % de caras y un 50 % de cruces: diremos que el porcentaje esperado converge a la probabilidad subyacente de salir cara. Por supuesto, al comienzo de la secuencia el ratio puede estar lejos del 50:50, por ejemplo porque tengamos una racha de caras, en cuyo caso estaremos tentados a pensar que a continuación «toca» cruz, para que los porcentajes se equilibren; esto se conoce como la «falacia del apostador» y es un sesgo psicológico que (lo digo desde la experiencia personal) es muy difícil de superar. Pero la moneda no tiene memoria; la idea con la que hay que quedarse es que la moneda no puede compensar por desequilibrios pasados, sino que simplemente los supera a base de más y más lanzamientos independientes.

En el capítulo 3 presentamos la clásica «curva con forma de campana», también conocida como distribución normal o gaussiana, la cual vimos que describía bien la distribución de los pesos al nacer en la población de Estados Unidos, y argumentamos que esto se debe a que el peso al nacer depende de un enorme número de factores, todos los cuales tienen una influencia limitada —cuando sumamos todos esos pequeños efectos obtenemos la distribución normal—.

Esta es la lógica detrás de lo que se conoce como teorema central del límite, demostrado por vez primera en 1733 por el matemático francés Abraham de Moivre para el caso particular de la distribución binomial. Pero no es solo la distribución binomial la que tiende a una curva normal con el incremento del tamaño de la muestra; es un hecho notable el que, para grandes tamaños muestrales y virtualmente cualquier forma de la distribución poblacional a partir de la cual haya sido extraída cada una de las medidas originales, su media habrá sido extraída de una curva normal.[121] La media en este caso es igual a la

media de la distribución original y la desviación típica tiene una relación simple con la desviación típica de la distribución original de la población y, como ya se ha mencionado, se conoce normalmente como error típico.[122]

Además de su trabajo sobre la sabiduría de las multitudes, la correlación, la regresión y casi todo lo demás, Francis Galton también consideraba una auténtica maravilla el que la distribución normal, conocida entonces como ley de frecuencia del error, surja de manera ordenada a partir de un caos aparente:

Apenas conozco nada tan capaz de atrapar la imaginación como la maravillosa forma de orden cósmico expresada por la «ley de frecuencia del error». Habría sido personificada y deificada por los griegos si la hubieran conocido. Reina con una serenidad y sencillez completas, en medio de la más salvaje de las confusiones. Cuanto mayor es la muchedumbre, y mayor la aparente anarquía, más perfecto es su poder. Es la ley suprema de la Sinrazón. Siempre que una gran muestra de elementos caóticos es llevada de la mano y dirigida al orden, se demuestra que una forma de regularidad insospechada y enormemente bella había estado siempre latente.

Tenía razón: es realmente una extraordinaria ley de la naturaleza.

¿Cómo ayuda esta teoría a establecer la fiabilidad de nuestros estimadores?

Toda esta teoría está bien para proporcionar información sobre distribuciones de estadísticos basados en datos extraídos de poblaciones conocidas, pero no es eso en lo que estamos interesados. Tenemos que encontrar una forma de revertir el proceso: en lugar de ir de poblaciones conocidas a muestras posibles, necesitamos ir de una muestra simple a una población posible. Este es el proceso de inferencia inductiva que vimos en el capítulo 3.

Supongamos que tengo una moneda, y le pregunto por la probabilidad de que salga cara. Usted me responde rápidamente «50:50», o algo parecido. Acto seguido lanzo la moneda, escondo el resultado antes de que alguno de los dos lo vea y le pregunto de nuevo por la probabilidad de que salga cara. Por mi experiencia, lo normal es que, después de pensarlo brevemente, responda usted a regañadientes «50:50». Después echo un vistazo rápido a la moneda, sin mostrársela, y repito la pregunta. Una vez más, si es usted como la mayoría, murmurará finalmente «50:50».

Este ejercicio sencillo muestra la existencia de dos tipos de incertidumbre muy diferentes: lo que se conoce como incertidumbre aleatoria antes de que lance la moneda —el «azar» de un evento impredecible— y la incertidumbre epistémica después de que haya lanzado la moneda —una expresión de nuestra ignorancia personal sobre un evento que es fijo pero desconocido—. Existe la misma diferencia entre un billete de lotería (en el que el resultado depende del azar) y una tarjeta que hay que rascar (donde el resultado ya está decidido, pero no sabes cuál es).

La estadística se usa cuando nos enfrentamos a una incertidumbre epistémica sobre alguna cantidad. Por ejemplo, llevamos a cabo una encuesta para saber el porcentaje real de una población que se considera religioso, o llevamos a cabo un ensayo clínico si no sabemos el efecto medio real de un medicamento. Como hemos visto, estas cantidades fijas pero desconocidas se denominan parámetros y normalmente se los identifica mediante una letra del alfabeto griego. Al igual que en mi ejemplo de lanzar una moneda, antes de que hagamos estos experimentos ya tenemos incertidumbre aleatoria sobre cuáles podrían ser los resultados, debido al muestreo aleatorio de individuos o a la asignación aleatoria de pacientes al medicamento o al placebo. Una vez que hemos hecho el estudio y obtenido los datos, usamos este modelo probabilístico para enfrentarnos a nuestra actual incertidumbre epistémica, de la misma manera que estaba usted dispuesto al final a decir «50:50» acerca de la moneda tapada.[123] De manera que la teoría de la probabilidad, que nos dice qué esperar en el futuro, se usa para decirnos qué podemos aprender de lo que hemos observado en el pasado. Esta es la (bastante notable, por cierto) base de la inferencia estadística.

El procedimiento para calcular el intervalo de incertidumbre alrededor de nuestro estimador, o un margen de error, se basa en esta idea fundamental. Hay tres estadios:

1. Usamos la teoría de la probabilidad para que nos dé, para cualquier parámetro poblacional específico, un intervalo en el que esperamos que se encuentre el estadístico observado con una probabilidad del 95 %. Son estos los intervalos de predicción al 95 %, como los que se muestran en el embudo de la figura 9.2.

2. Después observamos un estadístico particular.

3. Finalmente (y esta es la parte difícil) establecemos el rango de posibles parámetros poblacionales para los cuales nuestro estadístico se encuentra dentro del intervalo de predicción del 95 %. Esto es lo

que denominamos «intervalo de confianza del 95 %».

4. Este intervalo de confianza resultante recibe la etiqueta «95 %», dado que si se aplica repetidamente, el 95 % de esos intervalos deberían contener el valor real.[124]

¿Ha quedado todo claro? Si no es así, piense, por favor, que su sentimiento ha sido compartido por generaciones de perplejos estudiantes. En el glosario se muestran fórmulas específicas, pero los detalles son menos importantes que los principios fundamentales; un intervalo de confianza es el rango de parámetros de población para los cuales nuestro estadístico observado es una consecuencia plausible.

Calcular intervalos de confianza

La idea de los intervalos de confianza fue formalizada en la década de 1930 en el University College de Londres por Jerzy Neyman, un brillante matemático y estadístico polaco, y Egon Pearson, el hijo de Karl Pearson.[125] Se había estudiado durante décadas cómo derivar las necesarias distribuciones de probabilidad de coeficientes de correlación estimados y coeficientes de regresión. En los cursos académicos estándar de estadística se enseñaban los detalles matemáticos de estas distribuciones, que incluso había que derivar a partir de los primeros principios. Afortunadamente, los resultados de estos laboriosos esfuerzos están ahora encapsulados en el software estadístico, de manera que los investigadores pueden centrarse en las cuestiones importantes sin distraerse por fórmulas complejas.

Vimos en el capítulo 7 cómo puede usarse el bootstrap para obtener intervalos del 95 % para el gradiente de la regresión de Galton de la estatura de las hijas sobre la de las madres. Es mucho más sencillo obtener intervalos que se basen en la teoría de la probabilidad y sean proporcionados por el software estándar, y la tabla 9.1 muestra que los resultados obtenidos son muy similares. Los intervalos «exactos» basados en la teoría de la probabilidad requieren más supuestos que el enfoque basado en el bootstrap, y estrictamente hablando solo serían correctos si la distribución poblacional subyacente fuese normal. Pero el teorema central del límite significa que con un tamaño muestral tan grande es razonable asumir que nuestros estimadores tienen una distribución normal, y, por ello, los intervalos exactos serán aceptables.

	Gradiente de la regresión de los progenitores con respecto a la descendencia		
	Estimador	Error típico	Intervalo del 95 %
Exacto	0,33	0,0s	0,23 a 0,42
Bootstrap	0,33	0,06	0,22 a 0,44

Tabla 9.1. Estimadores de los coeficientes de regresión que resumen la relación entre las estaturas de las madres y de las hijas, con errores típicos exactos y derivados de bootstrap e intervalos de confianza de 95 % (el bootstrap se basa en 1.000 re-muestreos).

Lo normal es usar los intervalos del 95 %, que generalmente se fijan en más o menos dos errores típicos, aunque también se adoptan intervalos más estrechos (por ejemplo, 80 %) o más amplios (por ejemplo, 99 %).[126] La Oficina de Estadísticas Laborales de Estados Unidos usa intervalos del 90 % para el desempleo, mientras que la Oficina de Estadísticas Nacionales del Reino Unido usa el 95 %: es esencial dejar claro cuál se está usando.

Márgenes de error en encuestas

Cuando está claro que una determinada afirmación se basa en una encuesta, como por ejemplo una encuesta de opinión pública, es práctica normal señalar el margen de error. Las estadísticas de desempleo introducidas en el capítulo 7 tienen unos márgenes de error sorprendentemente grandes: el cambio estimado de 3.000 tiene un margen de error de ±77.000. Esto tiene a su vez un efecto material sobre la interpretación del número original: en este caso el margen de error revela que no podemos siquiera estar seguros de si el desempleo ha aumentado o disminuido.

Una regla sencilla es que si se está estimando el porcentaje de personas que prefiere, por ejemplo, el café frente al té para desayunar, y se pregunta a una muestra aleatoria de la población, entonces el margen de error (en %) es como máximo más o menos 100 dividido por la raíz cuadrada del tamaño de la muestra.[127] Así que para una encuesta de 1.000 personas (el estándar de la profesión), el margen de error es generalmente ±3 %:[128] si 400 de ellos dicen que prefieren café, y 600 de ellos dicen que prefieren té, entonces se podría hacer una estimación aproximada de que el porcentaje subyacente de personas en la población que prefieren café es del 40 ± 3 %, o entre un 37 % y un 43 %.

Por supuesto, esto solo es correcto si la empresa encuestadora

realmente ha utilizado una muestra aleatoria, si todo el mundo ha contestado, si todos tenían una opinión formada al respecto y si todos han dicho la verdad. De manera que, aunque podamos calcular márgenes de error, debemos recordar que solo se sostienen si nuestras asunciones son aproximadamente correctas. Pero ¿podemos basarnos en estas asunciones?

¿Debemos creer en los márgenes de error?

Antes de las elecciones generales del Reino Unido de junio de 2017, se publicaron numerosas encuestas de opinión sobre la intención de voto de alrededor de 1.000 entrevistados. Si se hubiese tratado de encuestas perfectamente aleatorias a las cuales los participantes hubiesen dado respuestas verdaderas, entonces el margen de error de cada una de ellas habría sido como mucho del ±3 %, y, por consiguiente, la variabilidad de las encuestas alrededor de su media habría estado en ese rango, dado que se suponía que todas medían la misma población subyacente. Pero la figura 9.3, basada en un gráfico empleado por la BBC, muestra que la variabilidad es mucho mayor que esto, lo que quiere decir que los márgenes de error no podían ser correctos.

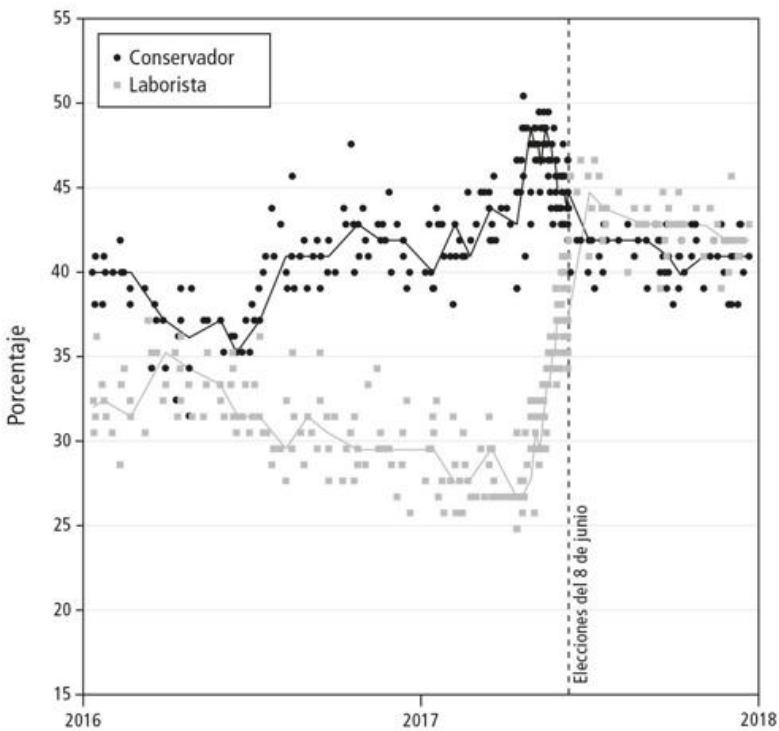


Figura 9.3. Estilo de visualización de los datos de opinión pública

usados por la BBC antes de las elecciones generales del Reino Unido del 8 de junio de 2017. (El gráfico de la BBC de las encuestas electorales está en: http://www.bbc.co.uk/news/election-2017-39856354). La línea de tendencia es la mediana de las anteriores siete encuestas. Cada encuesta estaba normalmente basada en 1.000 entrevistados, y por tanto afirmaba tener un margen de error de como máximo ±3 %. Pero la variabilidad entre encuestas excede este margen de largo. Solo se muestran el Partido Laborista y el Conservador.

Ya hemos visto muchas de las razones por las cuales las encuestas pueden ser incorrectas, más allá del inevitable (y cuantificable) margen de error debido a la variabilidad aleatoria. En este caso el exceso de variabilidad podría ser debido a los métodos de muestreo, en particular el uso de muestreo telefónico con una tasa de respuesta muy baja, quizás entre un 10 % y un 20 %, y principalmente usando líneas fijas. Mi heurística personal y más bien escéptica es que cualquier margen de error declarado en una encuesta debe ser multiplicado por dos para incluir los errores sistemáticos realizados en el muestreo.

Quizás no deberíamos esperar que las encuestas preelectorales sean de una precisión absoluta, pero sí que esperaríamos algo más de unos científicos que intentan medir hechos físicos como la velocidad de la luz. Pero, de hecho, hay toda una larga historia de márgenes de error declarados en experimentos de ese tipo que más tarde se han demostrado completamente inadecuados: en la primera mitad del siglo XX, los intervalos de incertidumbre alrededor de los estimadores de la velocidad de la luz no incluían los valores aceptados en la actualidad.

Esto ha llevado a organizaciones que trabajan en metrología, la ciencia de la medición, a especificar que los márgenes de error deberían basarse siempre en dos componentes:

→ Tipo A: la medida estadística estándar discutida en este capítulo, que se esperaría que se redujese a medida que aumentan las observaciones.

→ Tipo B: errores sistemáticos que no se esperaría que se redujesen con más observaciones, y tienen que ser manejados con medios no estadísticos como opinión de expertos o evidencia externa.

Todo esto nos debería llevar a ser humildes acerca de los métodos estadísticos que podemos aplicar a una única fuente de datos. Si hay problemas fundamentales con la forma en la que los datos han sido

reunidos, entonces no hay ningún método que pueda eliminar estos sesgos, y tendríamos que usar nuestro conocimiento y experiencia acerca del problema de fondo para moderar nuestras conclusiones.

¿Qué ocurre cuando tenemos todos los datos posibles?

Parece natural emplear la teoría de la probabilidad para calcular márgenes de error alrededor de resultados de encuesta, dado que los individuos han sido seleccionados aleatoriamente a partir de una población mayor, por lo que el azar tiene un papel claro en la producción de los datos. Pero de nuevo, nos hacemos la siguiente pregunta: ¿qué ocurre si la estadística mencionada se basa en todos los datos posibles? Por ejemplo, un país cuenta sus homicidios anualmente. Asumiendo que no hay error en el recuento (y que hay acuerdo acerca de qué significa «homicidio»), se trata simplemente de estadísticos descriptivos sin margen de error.

Pero supongamos que queremos decir algo sobre tendencias subyacentes a lo largo del tiempo, como que «la tasa de homicidios en el Reino Unido está aumentando». Por ejemplo, la Oficina de Estadísticas Nacionales del Reino Unido registró 497 homicidios entre abril de 2014 y marzo de 2015, y 557 el año siguiente. Es cierto que el número de homicidios ha aumentado, pero sabemos que el número de asesinatos varía de un año a otro sin razón aparente, así que ¿representa esto un cambio real en la tasa de homicidios subyacente anual? Queremos hacer una inferencia sobre esta cantidad desconocida, de manera que necesitamos un modelo probabilístico para nuestro recuento de homicidios observados.

Afortunadamente, vimos en el último capítulo que el número de homicidios diarios se comporta como si fueran observaciones aleatorias extraídas de una distribución Poisson derivada de una población metafórica de historias alternativas posibles. Esto a su vez significa que el total a lo largo de todo el año puede ser considerado una observación única de una distribución Poisson con media m igual a la (bastante hipotética) tasa anual subyacente «real». Lo que nos interesa es si m cambia de un año a otro.

La desviación típica de esta distribución Poisson es la raíz cuadrada de m, es decir, $\sqrt{m}$, que es también el error típico de nuestro estimador. Esto nos permitiría crear un intervalo de confianza, si supiésemos el valor de m. Pero no lo sabemos (ese es el objetivo de este ejercicio). Consideremos el período 2014-2015, cuando se produjeron 497 homicidios, que es nuestro estimador para la tasa subyacente m ese año. Podemos usar esta estimación de m para estimar el error típico

√m como = 22,3. Esto nos da un margen de error de ±1,96 × 22,3 = ±43,7. De manera que podemos obtener nuestro intervalo 95 % de m como 497 ± 43,7 = 453,3 a 540,7. Dado que se asume que los intervalos de confianza del 95 % son más o menos 1,96 errores típicos, esto significa que podemos confiar al 95 % en que la tasa de homicidios subyacente anual está entre 453 y 541.

La figura 9.4 muestra el número de homicidios observado en Inglaterra y Gales entre 1998 y 2016, con intervalos de confianza del 95 % para la tasa subyacente. Está claro que aunque hay una variación inevitable entre los recuentos anuales, los intervalos de confianza muestran que necesitamos ser cuidadosos al extraer conclusiones acerca de los cambios a lo largo del tiempo. Por ejemplo, el intervalo del 95 % alrededor de la cifra de 557 del período 2015-2016 iría de 511 a 603, con una superposición sustancial con el intervalo de confianza del año anterior.

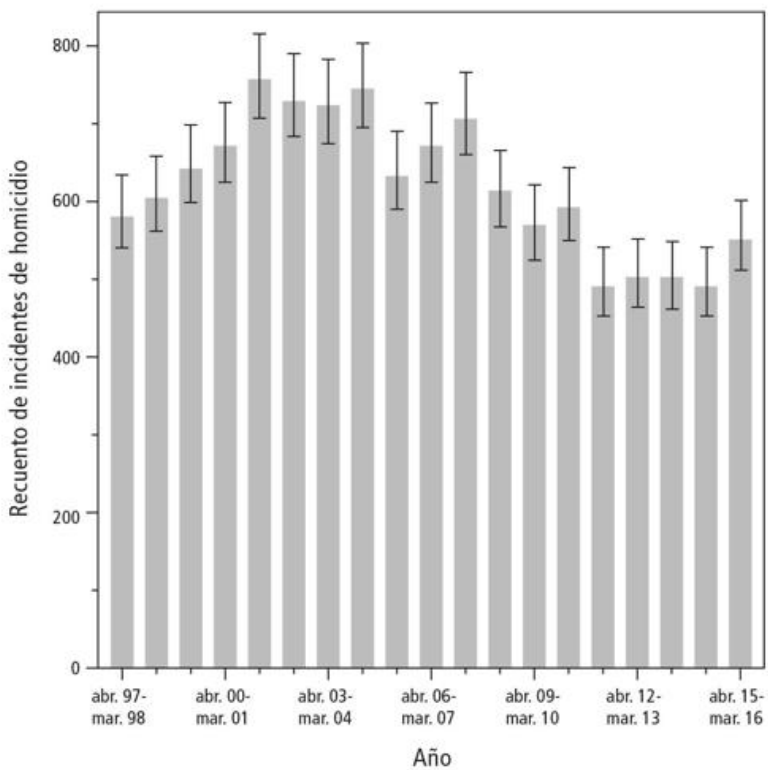


Figura 9.4. Número anual de homicidios en Inglaterra y Gales entre 1998 y 2016, e intervalos de confianza del 95 % para la tasa «real» subyacente de homicidios. (Márgenes de error para las estadísticas de homicidios: https://www.ons.gov.uk/

peoplepopulationandcommunity/crimeandjustice/compendium/focusonviolentcrimeandsexualoffences/yearendingmarch2016/homicide#statistical-interpretation-of-trend-in-homicides).

De manera que ¿cómo podemos decidir si ha habido un cambio real en el riesgo subyacente de ser una víctima de homicidio, o si los cambios observados pueden ser atribuidos a una variación aleatoria inevitable? Si los intervalos de confianza no se superponen, entonces podemos al menos estar un 95 % seguros de que se ha producido un cambio real. Pero este criterio es bastante estricto, por lo que deberíamos en realidad crear un intervalo del 95 % para el cambio en las tasas subyacentes. Y si este intervalo incluye el 0, no podemos estar seguros de que haya habido un cambio real.

Hubo un incremento de $557 - 497 = 60$ en el número de homicidios entre 2014-2015 y 2015-2016. Y resulta que un intervalo de confianza del 95 % alrededor de este cambio observado va de -4 a $+124$, que (por los pelos) incluye el 0. Técnicamente, esto significa que no podemos concluir con una confianza del 95 % que la tasa subyacente haya cambiado, pero dado que ese valor está justo en los márgenes sería poco razonable proclamar que no ha habido ningún cambio en absoluto.

Los intervalos de confianza alrededor de los recuentos de homicidios en la figura 9.4 son de una naturaleza completamente diferente a la de los márgenes de error en torno a, por ejemplo, las cifras del paro. Estas últimas son una expresión de nuestra incertidumbre epistémica sobre el número real de personas desempleadas, mientras que los intervalos alrededor de las cifras de homicidios no expresan incertidumbre sobre el número real de homicidios —asumimos que estos han sido correctamente registrados— sino sobre los riesgos subyacentes en la sociedad. Estos dos tipos de intervalos podrían parecer similares, e incluso usan unas matemáticas similares, pero tienen interpretaciones fundamentalmente diferentes.

El contenido de este capítulo ha sido un tanto complejo, lo cual no es sorprendente, dado que en él nos hemos ocupado esencialmente de todos los fundamentos formales de la inferencia estadística basada en los modelos de probabilidad. Pero el esfuerzo vale la pena, dado que ahora podemos usar esta base para ir más allá de las descripciones y estimaciones básicas de características del mundo, y empezar a ver cómo los modelos estadísticos nos pueden ayudar a responder cuestiones importantes sobre cómo funciona el mundo y proporcionar una base firme para los descubrimientos científicos.

Resumen

→ La teoría de la probabilidad se puede usar para derivar la distribución muestral de los estadísticos de resumen, de los cuales pueden derivarse a su vez fórmulas para los intervalos de confianza.

→ Un intervalo de confianza del 95 % es el resultado de un procedimiento que en el 95 % de los casos en los cuales sus asunciones son correctas, contendrá el valor real del parámetro. No se puede afirmar que un intervalo específico tiene una probabilidad del 95 % de contener el valor real.

→ El teorema central del límite implica que se puede asumir que las medias muestrales y otros estadísticos de resumen tienen una distribución normal en el caso de grandes muestras.

→ Los márgenes de error normalmente no incorporan error sistemático debido a causas no aleatorias; se requiere conocimiento externo y buen juicio para evaluarlos.

→ Los intervalos de confianza pueden ser calculados incluso cuando observamos todos los datos, en cuyo caso representan la incertidumbre sobre los parámetros de una población metafórica subyacente.

[119] Para obtener esta distribución, calculamos la probabilidad de dos zurdos como $0{,}2 \times 0{,}2 = 0{,}04$, la probabilidad de dos diestros como $0{,}8 \times 0{,}8 = 0{,}64$, por lo que la de uno de cada tiene que ser, entonces, $1 - 0{,}04 - 0{,}64 = 0{,}32$.

[120] El blog original de Paul está en https://pb204.blogspot.com/2011/10/funnel-plot-of-uk-bowel-cancer.html, y los datos pueden bajarse de http://pb204.blogspot.co.uk/2011/10/uploads.html.

[121] Hay excepciones importantes: algunas distribuciones tienen colas tan largas, «pesadas», que sus esperanzas y desviaciones típicas no existen, de manera que las medias no tienen dónde converger.

[122] Si podemos asumir que todas nuestras observaciones son independientes y provienen de la misma distribución poblacional, el error típico de su media es solo la desviación típica de la distribución poblacional dividida por la raíz cuadrada del tamaño muestral.

[123] Veremos en el capítulo 12 que a aquellos que usan estadística

bayesiana les vale el uso de probabilidades para la incertidumbre epistémica sobre parámetros.

[124] En un sentido estricto, el intervalo de confianza del 95 % no significa que haya una probabilidad del 95 % de que este intervalo particular contenga el valor real, aunque en la práctica la gente normalmente le atribuye esta interpretación incorrecta.

[125] Tuve el placer de conocer a ambos cuando ya eran muy mayores.

[126] Más concretamente, los intervalos de confianza del 95 % se fijan normalmente a más o menos 1,96 errores típicos, basándose en asumir una distribución muestral normal para el estadístico.

[127] El margen de error es $\pm 2\sqrt{p(1-p)/n}$, cuyo valor máximo de $\pm 1/\sqrt{n}$ se produce a $p = 0{,}5$. Por tanto, el margen de error es como máximo $\pm 1/\sqrt{n}$, sea cual sea el valor del verdadero porcentaje subyacente p.

[128] Con 1.000 participantes el margen de error (en %) es como máximo $\pm 100\sqrt{1000} = 3$ %. Las encuestas podrían tener diseños más complejos que la simple extracción de una muestra aleatoria de la población, pero los márgenes de error no se verán muy afectados.

10

Responder preguntas

y reivindicar descubrimientos

¿Nacen más niños que niñas?

John Arbuthnot, un doctor que asumió el puesto de médico de la reina Ana en 1705, se empeñó en responder a esta pregunta, y para ello examinó datos sobre bautismos en Londres durante los 82 años transcurridos entre 1629 y 1710. Sus resultados se muestran en la figura 10.1, en términos de lo que ahora se conoce como la razón de los sexos, que es el número de niños por cada 100 niñas.

Descubrió que cada año hubo más bautizos de niños que de niñas, con una razón de sexos de 107 y una variación entre 101 y 116 a lo largo del período. Pero Arbuthnot quería enunciar una ley más general, y por ello argumentó que si realmente no hubiese ninguna diferencia en las tasas subyacentes de nacimientos de niños y niñas, entonces cada año habría una probabilidad de 50:50 de que naciesen más niños que niñas, o más niñas que niños, igual que si lanzásemos una moneda.

Pero obtener un exceso de niños cada año sería el equivalente a lanzar una moneda sin trucar 82 veces seguidas y sacar cara siempre. La probabilidad de que eso ocurra es de ½82, que es un número ciertamente pequeño, con 24 ceros después de la coma. Si observamos esto en un experimento real, podríamos afirmar con seguridad que la moneda está descompensada. En el mismo sentido, Arbuthnot concluyó que había alguna fuerza que generaba más niños, para contrarrestar la mayor mortalidad de los hombres: «Para reparar esa pérdida, la naturaleza previsora, por disposición de su sabio Creador, trae a este mundo más varones que hembras; y lo hace en una proporción casi constante».[129]

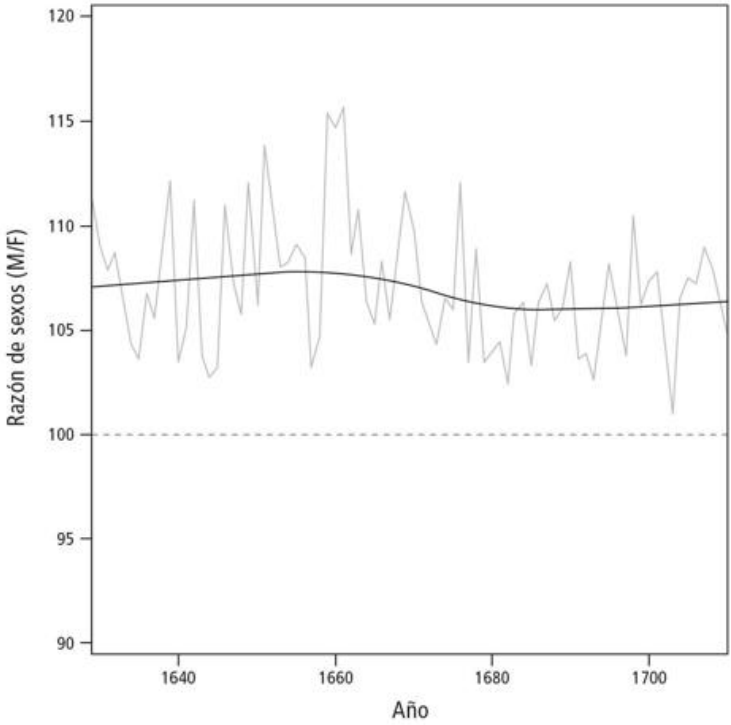


Figura 10.1. Razón de sexos (número de niños por cada cien niñas) para bautismos en Londres entre 1629 y 1710, a partir de la publicación de John Arbuthnot en 1710. La línea discontinua representa un número igual de niños y niñas; la curva está ajustada a los datos empíricos. En todos los años hubo más niños bautizados que niñas.

Los datos de Arbuthnot han sido sometidos a análisis repetidos, y aunque puede que haya algún error en las cuentas, y que solo se incluyan bautismos anglicanos, su descubrimiento básico se mantiene: ahora se considera que la razón de sexos «natural» está en torno a 105, lo que significa que por cada 20 niñas nacen 21 niños. El título del artículo que publicó usaba estos datos como una evidencia estadística directa de la existencia de una intervención sobrenatural: «Un Argumento en favor de la Providencia Divina, Derivado de la Regularidad Constante Observada en los Nacimientos de Ambos Sexos». Sea esta o no una conclusión justificada, y aunque no era consciente de ello en su época, ha entrado en la historia como el primero en realizar un test de significatividad estadística.

Hemos alcanzado quizás la parte más importante del ciclo de resolución de problemas, en el cual buscamos respuestas a cuestiones

específicas sobre cómo funciona el mundo. Por ejemplo:

1. ¿Sigue el número diario de homicidios en el Reino Unido la forma de una distribución Poisson?

2. ¿Ha cambiado la tasa de desempleo en el Reino Unido en el último trimestre?

3. ¿Tomar estatinas reduce el riesgo de infartos e ictus en gente como yo?

4. ¿Está la estatura de las madres asociada a la de sus hijos, una vez que tenemos en cuenta la estatura de los padres?

5. ¿Existe el bosón de Higgs?

Esta lista muestra que se pueden plantear preguntas muy diferentes, desde lo efímero a lo eterno:

1. Homicidios y la distribución Poisson: una regla general que no es de mucho interés público, pero ayuda a determinar si ha habido un cambio en las tasas subyacentes.

2. Cambios en la tasa de desempleo: una cuestión específica referida a un tiempo y lugar concretos.

3. Estatinas: una afirmación científica, pero referida a un grupo específico.

4. Estatura de las madres: posiblemente de interés científico general.

5. Bosón de Higgs: podría cambiar las ideas básicas que están detrás de las leyes físicas del universo.

Tenemos datos que nos pueden ayudar a resolver algunas de estas preguntas, con los cuales ya hemos llevado a cabo algunos gráficos exploratorios y extraído ciertas conclusiones informales sobre el modelo estadístico apropiado. Pero ahora nos enfrentamos a un aspecto formal de la parte de Análisis del ciclo PPDAC, conocido generalmente como comprobación de hipótesis.

¿Qué es una «hipótesis»?

Una hipótesis puede definirse como una explicación propuesta para un fenómeno. No es la verdad absoluta, sino más bien una asunción de trabajo, provisional, que quizá podría interpretarse como un sospechoso potencial en un caso penal.

Cuando discutimos la regresión en el capítulo 5, vimos que:

observación = modelo determinista + error residual

Esto refleja la idea de que los modelos estadísticos son representaciones matemáticas de lo que observamos, combinando un componente determinista con un componente «estocástico», este último referido a la impredictibilidad o «error» aleatorio, expresado generalmente en términos de una distribución de probabilidad. Dentro de la ciencia estadística, una hipótesis es considerada como una asunción particular sobre uno de estos componentes de un modelo estadístico, con la connotación de ser provisional, y no tanto «la verdad».

¿Por qué necesitamos comprobar formalmente las hipótesis nulas?

No son solo los científicos los que valoran los descubrimientos: el placer de encontrar algo nuevo es universal. De hecho, es algo tan deseable que hay una tendencia innata a pensar que hemos encontrado algo cuando no ha sido así. Hemos usado el término apofenia para describir la capacidad de ver pautas donde no las hay. En ocasiones se ha sugerido que esta tendencia podría incluso conferir una ventaja evolutiva: aquellos de nuestros ancestros que salían corriendo ante cualquier ruido proveniente de los arbustos sin esperar a ver si se trataba realmente de un tigre es más probable que sobreviviesen.

Pero aunque esta actitud puede haber sido adecuada para cazadores-recolectores, no funciona en el caso de la ciencia; de hecho, todo el proceso científico se iría al traste si sus postulados fueran simplemente fruto de nuestra imaginación. Tiene que haber una forma de protegernos contra descubrimientos falsos, y ese papel lo desempeña la comprobación de hipótesis.

La idea de la hipótesis nula es central en todo ello: es la forma simplificada de un modelo estadístico con la que vamos a operar hasta que tengamos suficiente evidencia en su contra. En las preguntas enumeradas más arriba, las hipótesis nulas serían las siguientes:

1. El número diario de homicidios en el Reino Unido sí que sigue una distribución Poisson.

2. La tasa de desempleo en el Reino Unido ha permanecido inalterada en el último trimestre.

3. Las estatinas no reducen el riesgo de infartos e ictus en personas

como yo.

4. La estatura de las madres no tiene ningún efecto en la estatura de los hijos, una vez que controlamos por la estatura de los padres.

5. El bosón de Higgs no existe.

La hipótesis nula es lo que asumimos que es la verdad hasta que se demuestra lo contrario. Es invariablemente negativa, niega todo progreso y cambio. Pero eso no significa que realmente creamos que la hipótesis nula es literalmente cierta: debe quedar claro que ninguna de las hipótesis que acabamos de enumerar puede ser, plausivamente, correcta (excepto posiblemente la inexistencia del bosón de Higgs). De manera que nunca podemos afirmar que la hipótesis nula ha sido realmente demostrada: en palabras de otro gran estadístico británico, Ronald Fisher, «la hipótesis nula nunca es demostrada o establecida, pero puede ser rechazada en el curso de la experimentación. Se puede decir que cada experimento existe solo para dar a los hechos una oportunidad de rechazar la hipótesis nula».[130]

Hay una marcada analogía con los juicios penales en el sistema legal inglés: un acusado puede ser declarado culpable, pero nunca nadie es declarado inocente, simplemente no se ha probado que sea culpable. Igualmente, encontraremos que podemos rechazar la hipótesis nula, pero si no tenemos suficiente evidencia al respecto no significa que podamos aceptarla como una verdad. Es solo una asunción de trabajo hasta que encontremos algo mejor.

Crúcese de brazos. ¿Ha puesto encima el brazo derecho o el izquierdo? Varios estudios han mostrado que alrededor de la mitad de la población pone su brazo derecho encima, y alrededor de la mitad, el izquierdo.

Pero ¿está esto asociado con ser hombre o mujer?

Aunque quizá no sea esta la pregunta científica más urgente, es algo que investigué cuando daba clases en el Instituto Africano de Ciencias Matemáticas en 2013 —fue un buen ejercicio de clase, y estaba genuinamente interesado en la respuesta—.[131] Obtuve datos de 54 estudiantes de posgrado originarios de toda África. La tabla 10.1 muestra las respuestas totales por género y si se ponía encima el brazo izquierdo o el derecho. Este tipo de tabla se conoce como tabla de contingencia.

En general, la mayoría colocaba encima su brazo derecho (32/54 = 59 %). Sin embargo, había un porcentaje mayor de mujeres (9/14 =

64 %) que de hombres (23/40 = 57 %) «brazoderechistas»: la diferencia observada en porcentaje es de 64 % − 57 % = 7 %. En este caso, la hipótesis nula sería que no hay realmente ninguna asociación entre el cruzarse de brazos y el género, en cuyo caso esperaríamos que la diferencia observada en los porcentajes entre géneros fuese del 0 %. Pero, por supuesto, la inevitable variabilidad aleatoria entre personas, incluso bajo esta asunción nula, significa que la diferencia observada no va a ser exactamente del 0 %. La cuestión crucial es si la diferencia observada del 7 % es lo suficientemente grande como para proporcionar evidencia en contra de la hipótesis nula.

	Mujer	Hombre	Total
Brazo izquierdo encima	5	17	22
Brazo derecho encima	9	23	32
Total	14	40	54

Tabla 10.1. Tabla de contingencia de los géneros y el comportamiento de cruzarse de brazos para 54 estudiantes de posgrado.

Para responder a esto, necesitamos saber qué tipo de diferencias observadas en los porcentajes esperaríamos encontrar simplemente por variación aleatoria; es decir, si la hipótesis nula fuese realmente cierta y el cruzarse de brazos fuese completamente independiente del género. Más formalmente, ¿es esta diferencia observada del 7 % compatible con la hipótesis nula?[132]

Esta es una idea algo compleja pero crucial. Cuando Arbuthnot estaba comprobando su hipótesis nula de que los niños y las niñas nacían con la misma probabilidad, podría haber deducido fácilmente que sus datos observados no eran en absoluto compatibles con esa asunción nula —la probabilidad de que los niños superasen a las niñas durante 82 años, si todo lo que estaba operando fuese el azar, era absolutamente minúscula—. En situaciones más complejas no está tan claro si los datos son compatibles con la hipótesis nula, pero el siguiente test de permutación es un potente procedimiento que no requiere de matemáticas complejas.

Imaginemos que ponemos a los 54 estudiantes en fila, 14 mujeres primero y 40 hombres después, y que a cada uno le asignamos un número del 1 al 54. Imaginemos también que cada uno de ellos tiene una etiqueta que identifica el brazo que coloca encima cuando se cruza de brazos. Ahora imaginemos que cogemos estas etiquetas de «cruce de brazos», las mezclamos en un sombrero y se las damos

aleatoriamente a los estudiantes. Este es un ejemplo de cómo esperaríamos que funcionase la naturaleza si la hipótesis nula fuera cierta, puesto que entonces el cruzar los brazos estaría completamente no relacionado con el género.

Incluso aunque el comportamiento de cruzar los brazos haya sido ahora asignado aleatoriamente, el porcentaje de brazoderechistas no será el mismo para mujeres que para hombres, por la operación del azar, y podemos calcular la diferencia observada en los porcentajes para este re-etiquetado aleatorio de los estudiantes. Luego podemos repetir este proceso de asignación aleatoria de comportamiento de cruce de brazos, digamos, un millar de veces, y ver qué distribución de diferencias se genera. Los resultados se muestran en la figura 10.2(a), un gráfico de dispersión de diferencias observadas —algunas a favor de los hombres, otras de las mujeres— centradas en una diferencia de cero. La diferencia observada real está cerca del centro de esta distribución.

Un enfoque alternativo, si tuviésemos mucho tiempo, sería llevar a cabo sistemáticamente todas las posibles permutaciones de las etiquetas sobre cruzado de brazos, y no quedarnos en 1.000 simulaciones. Cada una de ellas generaría una diferencia observada en los porcentajes de brazoderechismo entre hombres y mujeres, y hacer un gráfico con estas diferencias generaría una distribución más suave que la que generan solo 1.000 simulaciones.

Por desgracia, hay un enorme número de esas permutaciones, e incluso si las calculásemos a razón de un millón por segundo, el número de años que nos llevaría realizarlas tiene 57 ceros.[133] Afortunadamente, no tenemos que hacer esos cálculos, dado que la distribución de probabilidad para la diferencia observada en porcentajes de acuerdo con la hipótesis nula puede ser deducida en teoría, lo que se muestra en la figura 10.2(b); que se basa en lo que se conoce como la distribución hipergeométrica, que nos da la probabilidad de una celda concreta de la tabla considerando cada valor posible derivado de permutaciones aleatorias.

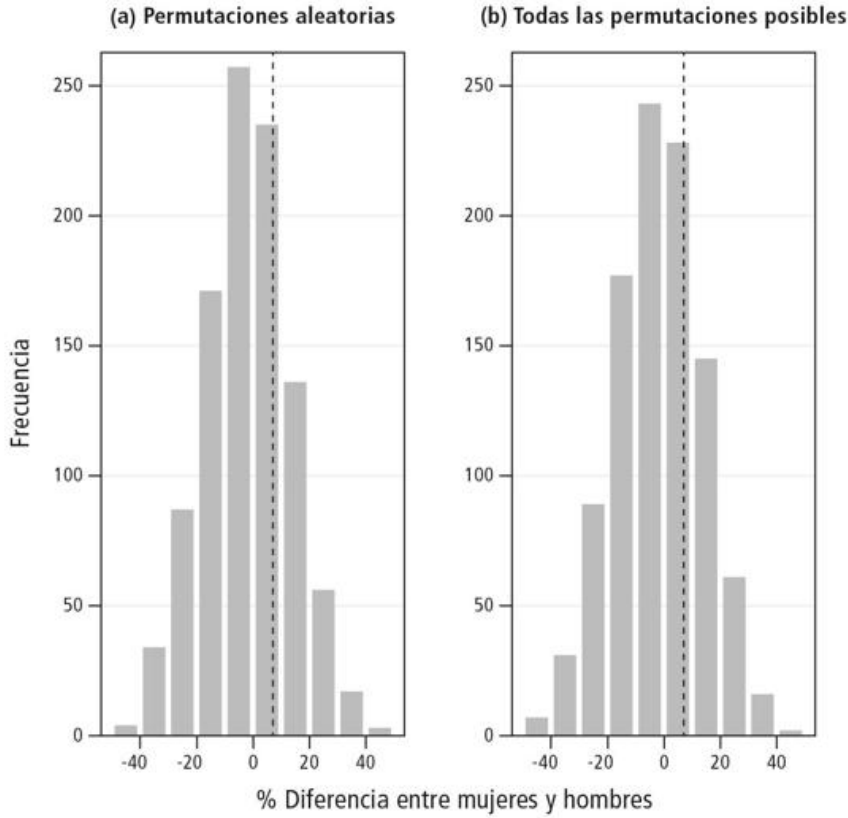


Figura 10.2. Distribución empírica de la diferencia en el porcentaje de mujeres y hombres que se cruzan de brazos con el brazo derecho encima: (a) a partir de 1.000 permutaciones aleatorias de cruce de brazos, (b) de todas las posibles permutaciones igualmente posibles de la respuesta de cruce de brazos. La diferencia observada en porcentajes (7 %) se indica con la línea vertical discontinua.

La figura 10.2 muestra que las diferencias observadas reales en el porcentaje de brazoderechistas (7 % en favor de las mujeres) están bastante cerca del centro de la distribución de las diferencias observadas que esperaríamos ver, si realmente no hubiese ningún tipo de asociación. Necesitamos una medida para resumir lo cercano al centro que está nuestro valor observado. Una medida de ese tipo es el «área de la cola» a la derecha de la línea discontinua mostrada en la figura 10.2, que es 45 %, o 0,45.

Esta área de la cola se conoce como el valor P, uno de los conceptos más importantes de la estadística practicada en la actualidad, y que, por ello, merece que incluyamos aquí una definición formal: un valor

P es la probabilidad de obtener un resultado al menos tan extremo como el que hemos obtenido, si la hipótesis nula (y todas las demás asunciones del modelo) fuera realmente cierta.

La cuestión, por supuesto, es qué queremos decir por «extremo». Nuestro valor P de 0,45 es de una cola, dado que solo mide la probabilidad de que hayamos observado ese valor extremo a favor de las mujeres, si la hipótesis nula fuese realmente cierta. Este valor P corresponde a lo que se conoce como prueba unilateral. Pero un porcentaje esperado a favor de los hombres también nos llevaría a sospechar que la hipótesis nula no se sostiene. Deberíamos, por tanto, calcular también la probabilidad de obtener una diferencia observada de al menos el 7 % en cualquiera de las direcciones. Esto se conoce como un valor P de dos colas, correspondiente a una prueba bilateral. Esta área total de ambas colas es de 0,89, y dado que este valor está cercano a 1, indica que el valor observado está cerca del centro de la distribución nula. Por supuesto, esto se puede ver al instante en la figura 10.2, pero no siempre podemos contar con gráficos de este tipo, por lo que necesitamos un número que resuma formalmente lo extremos que son nuestros datos.

Arbuthnot proporcionó el primer ejemplo que se conoce de este proceso: bajo la hipótesis nula de que la probabilidad de que nazcan niños y niñas es la misma, la probabilidad de que los niños superen a las niñas a lo largo de 82 años era de ½82. Esto solo define lo extremos que son los datos en términos del exceso de niños con respecto a las niñas, pero también dudaríamos de la hipótesis nula si las niñas superasen a los niños, por lo que deberíamos doblar ese número a ½81, para obtener la probabilidad de un resultado extremo en cada dirección. De manera que ½81 podría ser considerado el primer valor P conocido, aunque ese término no se aplicaría hasta doscientos cincuenta años más tarde.

Mi pequeña muestra, por cierto, indica que no hay ningún vínculo entre el género y el cruzado de brazos, y, de hecho, otros estudios más científicos no han encontrado ninguna relación entre el cruzado de brazos y el género, el ser zurdo o diestro o cualquier otra característica.

Significatividad estadística

La idea de la significatividad estadística es muy clara: si un valor P es lo suficientemente pequeño, entonces decimos que los resultados son estadísticamente significativos. Este término fue popularizado por Ronald Fisher en la década de 1920 y, a pesar de las críticas que

veremos más adelante, continúa desempeñando un papel muy importante en la estadística.

Ronald Fisher era un hombre extraordinario pero difícil. Era extraordinario porque fue una figura pionera en dos campos distintos: la genética y la estadística. Pero era también famoso por su fuerte temperamento y podía ser extremadamente crítico con cualquiera que pensase que cuestionaba sus ideas. Por otro lado, su apoyo a la eugenesia y sus críticas públicas a la evidencia sobre el vínculo entre el tabaquismo y el cáncer de pulmón perjudicaron su imagen. Su reputación personal se vio dañada una vez que se revelaron sus vínculos financieros con la industria del tabaco, pero su reputación científica se mantiene inmaculada, y sus ideas siguen teniendo múltiples aplicaciones nuevas en el análisis de grandes bases de datos.

Como mencioné en el capítulo 4, Fisher desarrolló la idea de la aleatorización en pruebas agrícolas cuando trabajaba en la Estación Experimental de Rothamsted. Posteriormente ilustró sus ideas sobre la aleatorización en diseños experimentales, con su famoso test de la cata del té, en el que una mujer (se supone que una tal Muriel Bristol) afirmaba ser capaz de determinar, con solo probar una taza de té, si la leche había sido añadida antes o después de que el té se hubiese servido en la taza.

Se prepararon cuatro tazas en las que la leche había sido añadida antes, y otras cuatro en las que el té se había añadido antes. Las ocho tazas se presentaban en un orden aleatorio; a Muriel se le dijo que había cuatro de cada, y tenía que adivinar las cuatro a las que se les había añadido primero la leche. Se dice que acertó. Aplicando la distribución hipergeométrica vemos que hay una probabilidad de 1 sobre 70 de que, siguiendo la hipótesis nula, estuviese simplemente haciendo una conjetura. Este es un ejemplo de un valor P que por convención sería considerado pequeño, de manera que los resultados se declararían como evidencia estadísticamente significativa de que ella sí podía, de hecho, decir si la leche se había añadido antes del té o no.

En resumen, he descrito los siguientes pasos:

1. Plantear una cuestión en términos de una hipótesis nula que queremos comprobar. A esto se le da normalmente la notación H0.

2. Escoger un test estadístico que estime algo que, si resulta ser suficientemente extremo, nos llevaría a dudar de la hipótesis nula (a menudo valores mayores del estadístico indican incompatibilidad con

la hipótesis nula).

3. Generar la distribución muestral de este test estadístico, en el caso de que la hipótesis nula sea cierta.

4. Comprobar si nuestros estadísticos observados están en las colas de esta distribución y resumir esto con el valor P: la probabilidad, si fuera cierta la hipótesis nula, de observar un estadístico tan extremo. El valor P es, por tanto, un área particular de la cola.

5. «Extremo» tiene que ser definido cuidadosamente: si tanto los valores grandes positivos como los negativos del test estadístico hubiesen sido considerados incompatibles con la hipótesis nula, el valor P tiene que tener esto en cuenta.

6. Declarar que los resultados son estadísticamente significativos si el valor P está por debajo de algún umbral crítico.

Ronald Fisher usó $P < 0{,}05$ y $P < 0{,}01$ como umbrales críticos convenientes para indicar significatividad, y produjo tablas de los valores críticos necesarios para alcanzar esos niveles de significatividad. La popularidad de esas tablas llevó a que 0,05 y 0,01 se convirtiesen en valores convencionales, aunque ahora se recomienda que se declaren los valores P exactos. Y es importante señalar que el valor P exacto es condicional no solo a la veracidad de la hipótesis nula, sino también a otros supuestos subyacentes al modelo estadístico, como la falta de sesgo sistemático, que las observaciones sean independientes y así sucesivamente.

Todo este proceso es conocido como test de significatividad de la hipótesis nula (TSHN), y, como veremos más adelante, es muy controvertido. Pero primero debemos examinar cómo se usan en la práctica las ideas de Fisher.

Usar la teoría de la probabilidad

Quizás el componente más complejo del test de significatividad de la hipótesis nula es el paso 3 —establecer la distribución del test estadístico escogido de acuerdo con la hipótesis nula—. Siempre podemos recurrir a métodos de simulación por ordenador, pero es mucho más conveniente usar la teoría de la probabilidad para establecer directamente las áreas de cola de los test estadísticos, como hizo Arbuthnot en su sencillo ejemplo, y Fisher con la distribución hipergeométrica.

A menudo usamos aproximaciones que fueron desarrolladas por los

pioneros de la inferencia estadística. Por ejemplo, alrededor de 1900 Karl Pearson desarrolló una serie de estadísticos para comprobar la asociación en tablas de contingencia como la de la tabla 10.1, de lo que surgió el test de asociación de chi-cuadrado.[134]

Estos test estadísticos calculan el número esperado de eventos en cada celda de la tabla, en el caso de que la hipótesis nula de no asociación sea cierta, y después un estadístico de chi-cuadrado mide la discrepancia total entre los recuentos observado y esperado. La tabla 10.2 muestra los números esperados en las celdas de la tabla, asumiendo la hipótesis nula: por ejemplo, el número esperado de mujeres que ponen el brazo izquierdo encima es el número total de mujeres (14), multiplicado por la proporción global de brazoizquierdistas (22/54), lo que da 5,7.

	Mujer	Hombre	Total
Brazo izquierdo encima	5 (5,7)	17 (16,3)	22
Brazo derecho encima	9 (8,3)	23 (23,7)	32
Total	14	40	54

Tabla 10.2. Recuento observado y esperado (entre paréntesis) de cruzado de brazos por género: los recuentos esperados se calculan bajo la hipótesis nula de que el cruzado de brazos no está asociado con el género.

Queda claro a partir de la tabla 10.2 que los recuentos observado y esperado son bastante similares, lo que refleja que los datos son lo que esperaríamos bajo la hipótesis nula. El estadístico de chi-cuadrado es una medida global de la disimilitud entre los recuentos observado y esperado (su fórmula se proporciona en el glosario), y tiene el valor 0,02. El valor P correspondiente a este estadístico, obtenible mediante cualquier software estándar, es de 0,90, lo que muestra que no hay evidencia contraria a la hipótesis nula. Es reconfortante que este valor P sea esencialmente el mismo que el del test «exacto» basado en la distribución hipergeométrica.

El desarrollo y el uso de los test estadísticos y de los valores P ha ocupado tradicionalmente una buena parte de los cursos estándar de Estadística, y eso, desgraciadamente, ha dado al campo la reputación de procurar simplemente escoger la fórmula adecuada y usar las tablas adecuadas. Aunque este libro intenta adoptar una perspectiva más amplia en la materia, es no obstante valioso revisitar los ejemplos que hemos discutido a lo largo del libro con respecto a su significatividad estadística.

1. ¿Sigue una distribución Poisson el número diario de homicidios en el Reino Unido?

La figura 8.5 mostró, para Inglaterra y Gales entre 2014 y 2016, el recuento observado de días con diferente número de homicidios. Hubo un total de 1.545 incidentes a lo largo de 1.095 días, una media de 1,41 al día, y bajo la hipótesis nula de una distribución Poisson con esta media, esperaríamos el recuento mostrado en la columna final de la tabla 10.3. Si adoptamos el enfoque usado para el análisis de la tabla 10.2, la discrepancia entre el recuento observado y esperado puede ser resumida por medio de un test de bondad de ajuste chi-cuadrado (una vez más, para más detalles, diríjase al glosario).

Número diario de incidentes de homicidio	Días observados	Días esperados según la hipótesis nula
0	259	267,1
1	387	376,8
2	261	265,9
3	131	125,0
4	40	44,1
5	13	12,4
6 o más	3	3,6
Total	1.095	1.095

Tabla 10.3. Días observados y esperados con un número específico de incidentes de homicidio en Inglaterra y Gales, de abril de 2014 a marzo de 2016. El test de bondad de ajuste chi-cuadrado tiene un valor P de 0,96, lo que indica que no hay evidencia en contra de la hipótesis nula de que la distribución es una Poisson.

El valor P observado de 0,96 no es significativo, por lo que no hay evidencia para rechazar la hipótesis nula (de hecho, el ajuste es tan bueno que es casi sospechoso). Por supuesto, de ahí no podemos asumir que la hipótesis nula sea exactamente cierta, pero debería ser razonable usarla como asunción cuando evaluamos, por ejemplo, los cambios en las tasas de homicidios que hemos visto en el capítulo 9.

2. ¿Ha cambiado recientemente la tasa de desempleo en el Reino Unido?

En el capítulo 7 vimos que un cambio trimestral de 3.000 en el desempleo tenía un margen de error de ± 77.000, basado en ± 2 errores típicos. Esto significa que el intervalo de confianza del 95 % va de -80.000 a $+74.000$ y claramente contiene el valor 0, que supone

que no ha habido cambios en el desempleo. Pero el hecho de que este intervalo del 95 % incluya el 0 es lógicamente equivalente a que el valor estimado (-3.000) sea menor que dos errores típicos desde 0, lo que quiere decir que el cambio no es significativamente distinto de 0.

Esto nos revela que los test de hipótesis y los intervalos de confianza son esencialmente semejantes:

→ Un valor P de dos colas es menor de 0,05 si el intervalo de confianza del 95 % no incluye la hipótesis nula (generalmente 0).

→ Un intervalo de confianza del 95 % es el conjunto de hipótesis nulas que no son rechazadas con un valor $P < 0,05$.

Este vínculo íntimo entre test de hipótesis e intervalos de confianza debería ser bastante para que la gente dejase de interpretar erróneamente resultados que no son, desde el punto de vista estadístico, significativamente distintos de 0; esto no significa que la hipótesis nula sea realmente cierta, sino simplemente que un intervalo de confianza para el valor real incluye el 0. Desgraciadamente, como veremos más adelante, esta lección es muchas veces ignorada.

3. ¿Tomar estatinas reduce el riesgo de infartos e ictus en gente como yo?

La tabla 10.4 repite los resultados del Estudio de Protección Cardiaca (HPS) de la tabla 4.1, pero añade varias columnas que muestran la confianza con la que podemos tomarnos los beneficios de la estatina. Hay una fuerte conexión entre los errores típicos, los intervalos de confianza y los valores P. Los intervalos de confianza para la reducción del riesgo son aproximadamente ± 2 errores típicos (nótese que el HPS redondea las reducciones relativas). Los intervalos de confianza excluyen de largo la hipótesis nula de 0 %, equivalente a que la estatina no tenga ningún efecto, y, por consiguiente, los valores P son muy pequeños; de hecho, el valor P para la reducción de un 27 % en infartos es alrededor de 1 en 3 millones. Esta es la consecuencia de llevar a cabo un estudio tan masivo.

Evento	Porcentaje para 10.269 personas que recibieron estatina	Porcentaje para 10.267 personas que recibieron el placebo	% de reducción (relativa) del riesgo para aquellos que recibieron estatina	Error típico de reducción del riesgo	Intervalo de confianza para el % de reducción	Valor P
Infarto	8,7	11,8	27 %	4 %	21 % a 33 %	P < 0,0001
Ictus	4,3	5,7	25 %	5 %	15 % a 34 %	P < 0,0001
Muerte por cualquier causa	12,9	14,7	13 %	4 %	6 % a 19 %	P = 0,0003

Tabla 10.4. Resultados del Estudio de Protección Cardiaca, que muestran los efectos estimados relativos, sus errores típicos, intervalos de confianza y valores P que comprueban la hipótesis nula de «no efecto».

Se podrían usar otros estadísticos de resumen, como la diferencia en riesgos absolutos, pero todos deberían proporcionar valores P similares. Los investigadores del HPS se centran en la reducción proporcional, dado que esta es bastante constante entre subgrupos, y por tanto supone una medida única de resumen bastante buena. Hay varias formas distintas de calcular los intervalos de confianza, pero debería haber pocas diferencias entre ellas.

4. ¿Está asociada la estatura de las madres a la de sus hijos, una vez que controlamos por la estatura de los padres?

En el capítulo 5 mostramos una regresión lineal en la que la variable de respuesta (dependiente) era la estatura del hijo, y las variables explicativas (independientes) eran la estatura de la madre y del padre. Los coeficientes se mostraban en la tabla 5.3, pero sin tener en cuenta si podían ser considerados significativamente distintos de 0. Para ilustrar la forma en la que estos resultados aparecen en el software estadístico, la tabla 10.5 reproduce el resultado tal como aparece en un programa popular y gratuito, R.

	Estimador	Error típico	Valor-*t*	*Pr* (>|*t*|)
(Constante)	69,22882	0,10664	649,168	< 2 e–16 ***
Estatura de la madre	0,33355	0,04600	7,252	1,74 e–12 ***
Estatura del padre	0,41175	0,04668	8,820	< 2 e–16 ***

Tabla 10.5. Una reproducción del resultado en R de una regresión múltiple usando los datos de Galton, con la estatura del hijo como variable de respuesta, y las estaturas de la madre y del padre como variables explicativas. El valor-t es el estimador dividido por el error típico. La columna encabezada por Pr ($>|t|$) representa un valor P de dos colas; la probabilidad de obtener un valor-t tan grande, positivo o negativo, bajo la hipótesis nula de que la relación real es 0. La notación «2 e−16» significa que el valor P es menor de 0,0000000000000002 (son 15 ceros). La línea final muestra las interpretaciones de los asteriscos en términos de los valores P.

Como en la tabla 5.3, la constante es la media de las estaturas de los hijos, y los coeficientes (denominados «estimadores» en el resultado) representan el cambio esperado en estatura por cada diferencia de una pulgada de su madre y padre con respecto a la media de la estatura de madres y padres. El error típico se calcula a partir de una fórmula bien conocida, y es claramente pequeño con relación al tamaño de los coeficientes.

El valor-t, conocido como estadístico-t, es muy relevante, dado que es el vínculo que nos dice si la asociación entre una variable explicativa y la respuesta es estadísticamente significativa. El valor-t es un caso especial de lo que se conoce como la t de Student. «Student» era el seudónimo de William Gosset, que desarrolló el método en 1908 mientras estaba en comisión de servicio en el University College de Londres proveniente de la cervecería Guinness de Dublín. El valor-t es simplemente el estimadorerror típico (esto se puede comprobar en la tabla 10.5), por lo que se puede interpretar como lo lejos que está el estimador de 0, medido en número de errores típicos. Dado un valor-t determinado, y el tamaño muestral, el software puede proporcionar un valor P preciso; para muestras grandes, valores-t mayores de 2 o menores de -2 corresponden a $P < 0{,}05$, aunque estos umbrales serán mayores para tamaños muestrales más pequeños. R usa un sistema de estrellas para los valores P, desde una *, que indica P < 0,05, hasta tres estrellas* **, que indica $P < 0{,}001$. En la tabla 10.5, los valores-t son tan grandes que los valores P son minúsculos.

En el capítulo 6 vimos que un algoritmo podía ganar una competición de predicción por un margen muy pequeño. Cuando se predice el conjunto de validación para la supervivencia en el caso del Titanic, por ejemplo, un sencillo árbol de clasificación lograba la mejor puntuación de Brier (el cuadrado de la media del error de predicción), 0,139, solo ligeramente inferior a la puntuación de 0,142 obtenida por la red neuronal media (véase la tabla 6.4). Es razonable preguntarse si este pequeño margen de ventaja de $-0{,}003$ es estadísticamente

significativo, en el sentido de si puede o no ser explicado por variación aleatoria.

Esto es fácil de comprobar. El estadístico-t es −0,54, con un valor P de dos colas de 0,59.[135] ¡No hay, por tanto, una evidencia sólida de que el árbol de clasificación sea realmente el mejor algoritmo! Este tipo de análisis no es rutinario en las competiciones del tipo Kaggle, pero parece importante reconocer que el estatus de ganador depende de la selección aleatoria de casos en el conjunto de validación.

Los investigadores se pasan la vida analizando el tipo de resultado que se muestra en la tabla 10.5, a la espera de ver los centelleantes asteriscos indicadores de un resultado significativo que puedan plasmar en su siguiente artículo científico. Pero, como veremos a continuación, este tipo de búsqueda obsesiva de significatividad estadística puede llevar fácilmente a falsas ilusiones.

El peligro de llevar a cabo muchas pruebas de significatividad

Los umbrales habituales para declarar que algo es «significativo», $P < 0,05$ y $P < 0,01$, fueron seleccionados de manera bastante arbitraria por Ronald Fisher para sus tablas, en la época en que el cálculo exacto de valores P no era posible con las calculadoras mecánicas y eléctricas disponibles entonces. Pero ¿qué ocurre cuando llevamos a cabo muchas pruebas de significatividad, intentando en cada una de ellas conseguir que nuestro valor P sea menor de 0,05?

Supongamos que un medicamento realmente no funciona; que la hipótesis nula es cierta. Si hacemos un ensayo clínico, declararemos que el resultado es estadísticamente significativo si el valor P es menor de 0,05, y dado que el medicamento no es efectivo, la probabilidad de que esto ocurra es de 0,05, o 5 % —la definición misma de un valor P —. Esto sería considerado un falso positivo, dado que nos hace creer incorrectamente que el medicamento es efectivo. Si hacemos dos ensayos y nos fijamos en el resultado más extremo, la probabilidad de obtener al menos un resultado positivo —y, por tanto, un falso positivo— es cercana a 0,10 o 10 %.[136] La probabilidad de obtener al menos un falso positivo se incrementa rápidamente a medida que hacemos más ensayos; si hacemos diez ensayos de medicamentos inútiles la probabilidad de obtener al menos un resultado significativo al $P < 0,05$ es de hasta un 40 %. Esto se conoce como el problema de las pruebas múltiples, y se produce cuando se llevan a cabo muchas pruebas de significatividad y se declaran los resultados más significativos.

Un problema especial tiene lugar cuando los investigadores dividen los datos en muchos subconjuntos, realizan una prueba de hipótesis en cada uno de ellos y después se fijan en la que es más significativa. Un ejemplo clásico es un experimento llevado a cabo por prestigiosos investigadores en 2009 que implicaba someter a un sujeto a una serie de fotografías de personas que expresan distintas emociones, y luego realizar una resonancia magnética funcional para ver qué regiones del cerebro del sujeto mostraban una respuesta significativa, con $P < 0{,}001$.

Lo curioso de esta historia es que el «sujeto» era un salmón del Atlántico de cuatro libras que «no estaba vivo cuando se le hizo el escáner». De un total de 8.064 áreas en el cerebro de este gran pez muerto, 16 mostraron una respuesta significativa a las fotografías. En lugar de concluir que el salmón muerto tenía unas facultades milagrosas, el equipo investigador identificó correctamente el problema de las pruebas múltiples —8.000 pruebas de significatividad generarán obligatoriamente falsos positivos—.[137] Incluso si se usa un criterio estricto de $P < 0{,}001$, esperaríamos obtener 8 resultados significativos solo por azar.

Una forma de evitar este problema es exigir un valor P muy pequeño para considerar que algo es significativo, y el método más sencillo, conocido como la corrección de Bonferroni, es usar un umbral de $0{,}05 / n$, donde n es el número de pruebas realizadas. Así, por ejemplo, las pruebas en cada área del cerebro del salmón podrían ser llevadas a cabo demandando un valor P de $0{,}05 / 8.000 = 0{,}00000625$, o 1 entre 160.000. Esta técnica se ha convertido en la regla que seguir cuando se investiga la conexión del genoma humano con áreas asociadas con enfermedades: dado que hay aproximadamente 1.000.000 áreas para los genes, se exige un valor P menor de $0{,}05 / 1.000.000 = 1$ entre 20 millones para poder declarar que se ha hecho un descubrimiento.

De manera que cuando se están estimando un gran número de hipótesis al mismo tiempo, como en los escáneres cerebrales o la genómica, se puede emplear el método de Bonferroni para decidir si los resultados más extremos son significativos. También se han desarrollado técnicas sencillas que relajan ligeramente el criterio de Bonferroni para el segundo resultado más extremo, el tercero más extremo, y así sucesivamente, diseñadas para controlar el porcentaje global de «descubrimientos» que resultan ser afirmaciones falsas —la llamada tasa de descubrimientos falsos—.

Otra forma de evitar falsos positivos es exigir que se replique el

estudio original, realizando el experimento repetido en circunstancias enteramente diferentes, pero esencialmente con el mismo protocolo. Para nuevos productos farmacéuticos que se someten a la aprobación de la Administración de Alimentos y Medicamentos de Estados Unidos, el procedimiento estándar es llevar a cabo dos ensayos clínicos independientes, cada uno de los cuales debe mostrar beneficios clínicos que sean significativos a $P < 0{,}05$. Esto significa que la probabilidad global de aprobar un medicamento que no genere ningún beneficio es $0{,}05 \times 0{,}05 = 0{,}0025$ o, lo que es lo mismo, 1 entre 400.

5. ¿Existe el bosón de Higgs?

A lo largo del siglo XX, los físicos desarrollaron un «modelo estándar» con la intención de explicar las fuerzas que operan a nivel subatómico. Una pieza del modelo seguía estando sin demostrar: el «campo de Higgs» de energía que permea el universo, y confiere masa a partículas como los electrones a través de su propia partícula fundamental, el llamado bosón de Higgs. Cuando los investigadores del CERN finalmente declararon haber descubierto el bosón de Higgs en 2012, se anunció como un resultado «cinco sigma».[138] Pero pocos se habrían dado cuenta de que esto no era sino una forma de expresar significatividad estadística.

Cuando los investigadores hicieron un gráfico con la tasa en la cual ocurren eventos específicos para distintos niveles de energía, vieron que la curva tenía una «joroba» ahí donde se esperaría que existiese el bosón de Higgs. Resultó igualmente importante que una forma de bondad de ajuste chi-cuadrado revelase un valor P de menos de 1 en 3,5 millones, bajo la hipótesis nula de que Higgs no existía y la «joroba» era simplemente el resultado de variación aleatoria. Pero ¿por qué se dijo que era un descubrimiento «cinco sigma»?

Es normal en la física teórica informar de descubrimientos en términos de «sigmas». Un resultado «dos sigma» es una observación que está a dos errores típicos de la hipótesis nula (recordemos que usamos la letra griega sigma (σ) para representar la desviación típica de una población): las «sigmas» en la física teórica corresponden precisamente al valor-t en la salida de ordenador mostrada en la tabla 10.5 para el ejemplo de la regresión múltiple. Dado que una observación que nos da un valor P de dos colas de 1 en 3,5 millones —derivada de la prueba de chi-cuadrado— estaría a cinco errores típicos de la hipótesis nula, se dice que el bosón de Higgs tiene un resultado cinco sigma.

Claramente, el equipo del CERN no quería anunciar su «descubrimiento» hasta que el valor P fuese extremadamente pequeño. En primer lugar, necesitaban asegurarse de que las pruebas de significatividad hubiesen sido llevadas a cabo para todos los niveles de energía, no solo el de la prueba final de chi-cuadrado —este ajuste para pruebas múltiples se conoce en física como el «efecto de mirar en otro sitio»—. Pero principalmente lo que querían es estar seguros de que cada intento de replicar resultados llevase a la misma conclusión. Sencillamente, realizar una afirmación incorrecta sobre las leyes de la física sería demasiado embarazoso.

Para responder a la pregunta planteada al comienzo de esta sección: parece razonable asumir que el bosón de Higgs existe. Esta es la nueva

hipótesis nula hasta que, quizás, se sugiera una nueva teoría.

Teoría Neyman-Pearson

¿Por qué el Estudio de Protección Cardiaca necesita alrededor de 20.000 participantes?

El Estudio de Protección Cardiaca era enorme, y no por casualidad. Cuando planeaban el ensayo clínico, los investigadores tenían que decir cuántas personas se necesitaba asignar de manera aleatoria a los tratamientos con y sin estatinas. Todo ello tenía que tener unas firmes bases estadísticas para justificar el gasto que suponía un experimento de ese tipo. Su plan se basaba en ideas estadísticas que fueron desarrolladas por Jerzy Neyman y Egon Pearson, con los que ya nos hemos encontrado anteriormente como creadores de los intervalos de confianza.

La idea de los valores P y las pruebas de significatividad fue desarrollada por Ronald Fisher en la década de 1920 como forma de evaluar la idoneidad de una hipótesis determinada. Si se observa un valor P pequeño, entonces o bien ha ocurrido algo muy sorprendente, o la hipótesis nula no es cierta: cuanto menor es el valor P, más evidente resulta que la hipótesis nula puede ser una asunción inapropiada. La intención inicial es que fuese un procedimiento bastante informal, pero en la década de 1930 Neyman y Pearson desarrollaron una teoría de comportamiento inductivo que intentaba aportar una fundamentación matemática más rigurosa a las pruebas de hipótesis.

Para ello necesitaban especificar no solo la hipótesis nula, sino también una hipótesis alternativa que representase una explicación más compleja de los datos. Consideraron entonces la decisión posible después de realizar una prueba de hipótesis, que sería o bien rechazar la hipótesis nula en favor de la alternativa, o no rechazar la hipótesis nula.[139] De manera que hay dos tipos de errores posibles: un error de tipo I se comete cuando rechazamos una hipótesis nula que es cierta, mientras que un error de tipo II se comete cuando no rechazamos la hipótesis nula y realmente la que se sostiene es la hipótesis alternativa. Hay una clara analogía jurídica, ilustrada en la tabla 10.6: un error legal de tipo I sería condenar falsamente a un inocente, y un error de tipo II sería declarar a alguien «no culpable» cuando realmente ha cometido el delito.

Cuando se diseña un experimento, la sugerencia de Neyman y Pearson es que deberíamos escoger dos cantidades que determinen

conjuntamente el tamaño del mismo. En primer lugar, deberíamos fijar la probabilidad de un error de tipo I, dado que la hipótesis nula es cierta, a un valor preespecificado, digamos, 0,05; esto se conoce como nivel de significatividad, también llamado α (alfa). En segundo lugar, debemos preespecificar la probabilidad de un error de tipo II, dado que la hipótesis alternativa es cierta, conocido generalmente como β (beta). De hecho, los investigadores normalmente emplean 1 − β, lo que se denomina potencia, la probabilidad de rechazar la hipótesis nula en favor de la alternativa, dado que esta última es cierta. En otras palabras, el poder de un experimento es la probabilidad de detectar correctamente un efecto real.

Verdad	Resultado de la prueba de hipótesis	
	No se rechaza la hipótesis nula (el sospechoso es declarado «no culpable»).	Se rechaza la hipótesis nula en favor de la alternativa (el sospechoso es declarado culpable).
Hipótesis nula (el sospechoso es inocente)	Se acierta en no rechazar la hipótesis nula. Se acierta en declarar a un inocente «no culpable».	Error de tipo I: rechazar incorrectamente la hipótesis nula. Condenar erróneamente a un inocente.
Hipótesis alternativa (el sospechoso es culpable)	Error de tipo II: no rechazar incorrectamente la hipótesis nula. No condenar a una persona culpable.	Rechazar correctamente la hipótesis nula. Condenar correctamente a una persona culpable.

Tabla 10.6. Resultados posibles de una prueba de hipótesis, utilizando la analogía de un juicio penal.

Hay una conexión muy estrecha entre el tamaño de α y el valor P de Fisher. Si tomamos α como el umbral a partir del cual consideramos que los resultados son significativos, entonces los resultados que nos llevan a rechazar la hipótesis nula serán exactamente aquellos para los que P es menor que α. De manera que α puede ser considerada como el nivel mínimo de significatividad —una α de 0,05 significa que rechazamos la hipótesis nula para todos los valores P menores de 0,05 — .

Existen fórmulas para la significatividad y la potencia de distintos tipos de experimentos, y cada una de ellas depende del tamaño muestral. Pero si el tamaño muestral es fijo, tenemos que llegar a un compromiso: para incrementar la potencia, siempre podemos hacer menos estricto el umbral de «significatividad», y, por tanto, hacer más probable que identifiquemos correctamente un efecto real, pero esto significa incrementar la probabilidad de un error de tipo I. Volviendo a la analogía jurídica, podemos relajar los requisitos necesarios para una condena, por ejemplo relajando el requisito de la prueba «más allá

de toda duda razonable», lo que resultará en que más delincuentes sean correctamente condenados, pero al coste inevitable de que más personas inocentes serán incorrectamente declaradas culpables.

La teoría Neyman-Pearson tenía sus raíces en el control de calidad industrial, pero ahora se usa ampliamente en la comprobación de nuevos tratamientos médicos. Antes de empezar a aleatorizar ensayos clínicos, el protocolo especificará una hipótesis nula según la cual el tratamiento no tiene efectos, y una hipótesis alternativa, que consistirá generalmente en un efecto que se considere al mismo tiempo plausible e importante. Después, los investigadores establecerán la significatividad y la potencia del estudio, normalmente $\alpha = 0{,}05$ y $\beta = 0{,}80$. Esto significa que exigirán un valor P menor de 0,05 para declarar que el resultado es significativo, con una probabilidad del 80 % de lograrlo si el tratamiento es realmente efectivo. De esta manera, se obtiene una estimación del número de participantes necesarios.

Los investigadores tienen que ser más estrictos si quieren llevar a cabo un ensayo clínico que resulte definitivo. Por ejemplo, el Estudio de Protección de la Salud concluyó que:

Si la terapia de reducción del colesterol reducía la mortalidad por enfermedad coronaria en aproximadamente un 25 % y la mortalidad por cualquier causa en aproximadamente un 15 %, entonces un estudio de este tamaño con una alta tasa de participación tendría grandes posibilidades de demostrar esos efectos con unos niveles convincentes de significatividad estadística (<90 % de potencia para lograr $p < 0{,}01$).

En otras palabras, si el efecto real del tratamiento es una reducción del 25 % en la mortalidad por enfermedad cardiaca y un 15 % en la mortalidad por cualquier causa (la hipótesis alternativa), el estudio tiene una potencia de alrededor de $\beta = 90$ % y $\alpha = 1$ % de significatividad. Estos requisitos requieren un tamaño muestral de alrededor de 20.000. De hecho, como muestra la tabla 10.4, los resultados finales incluían un 13 % de reducción de la mortalidad por todas las causas, lo que está notablemente cerca de lo que se había planeado.

La idea de tener una muestra suficientemente grande para tener potencia bastante para detectar una hipótesis alternativa plausible se ha convertido en un criterio fundamental en la planificación de los estudios médicos. Pero los estudios en psicología y neurociencia muchas veces tienen tamaños muestrales escogidos en base a la conveniencia o la tradición, y pueden ser incluso de solo 20 sujetos

por condición. Es cierto que podrían pasarse por alto hipótesis alternativas si los estudios son demasiado pequeños. Por ello, está siendo reconocida finalmente la necesidad de que otras áreas experimentales tengan en cuenta la potencia de sus experimentos.

Como veremos en el capítulo siguiente, Neyman y Pearson discutieron de manera vehemente, incluso agresiva, con Fisher sobre la forma apropiada de la prueba de hipótesis, un conflicto que no ha llevado a identificar cuál es el enfoque «correcto». El Estudio de Protección Cardiaca muestra que los ensayos clínicos tienden a ser diseñados desde la perspectiva Neyman-Pearson, pero, para ser exactos, la significatividad y la potencia son irrelevantes una vez que se ha llevado a cabo el experimento. En ese momento, el ensayo es analizado empleando intervalos de confianza para mostrar los valores plausibles para los efectos del tratamiento, y valores P fisherianos para resumir la solidez de la evidencia en contra de la hipótesis nula. De manera que una extraña mezcla de las ideas de Fisher y de Neyman-Pearson ha resultado ser notablemente efectiva.

¿Podría haberse atrapado antes a Harold Shipman?

Vimos en la introducción que el doctor Harold Shipman había asesinado a unos doscientos de sus pacientes durante veinte años antes de ser finalmente detenido. Las familias de sus víctimas se mostraron comprensiblemente afligidas ante el hecho de que hubiese llevado a cabo sus crímenes durante tanto tiempo sin levantar sospechas. Por ello, se puso en marcha una investigación pública para determinar si podía haber sido identificado con anterioridad. Antes de que comenzase la investigación propiamente dicha, se reunieron los certificados de fallecimiento firmados por Shipman referidos a personas que habían muerto en sus casas o en su clínica desde 1977, y se comparó con el número de personas que habrían fallecido, dada la composición de edad de todos los pacientes al «cuidado» de Shipman y las tasas de mortalidad de otros médicos de familia en el área circundante. Para hacer este tipo de comparación hay que controlar por las condiciones locales, como por ejemplo el cambio de temperatura y los brotes de gripe. La figura 10.3 muestra los resultados derivados de sustraer el número esperado del número observado de certificados de defunción, acumulados entre 1977 y la detención de Shipman en 1998. Esta diferencia podría considerarse su «exceso» de mortalidad.

Hacia 1998, su exceso de mortalidad para mayores de 65 años fue de 174 mujeres y 49 hombres. Se trata de casi exactamente el número de personas mayores que fueron confirmadas como víctimas por la

investigación pública, lo que muestra la notable exactitud de este análisis puramente estadístico, que no incluía información de casos individuales.[140]

Imaginémonos una historia ficticia en la que alguien hubiese estado siguiendo las muertes de Shipman año tras año y haciendo los cálculos necesarios para producir la figura 10.3. ¿En qué momento podría haber dado la voz de alarma? Se podría, por ejemplo, haber realizado una prueba de significatividad al final de cada año. Los recuentos de muertes, como los homicidios, son el resultado de muchos individuos con una pequeña probabilidad de que les suceda el evento. Se puede asumir que tendrá una distribución Poisson, por lo que la hipótesis nula será que el número observado acumulado de muertes es una observación de una distribución Poisson con una expectativa dada por el recuento acumulado esperado.

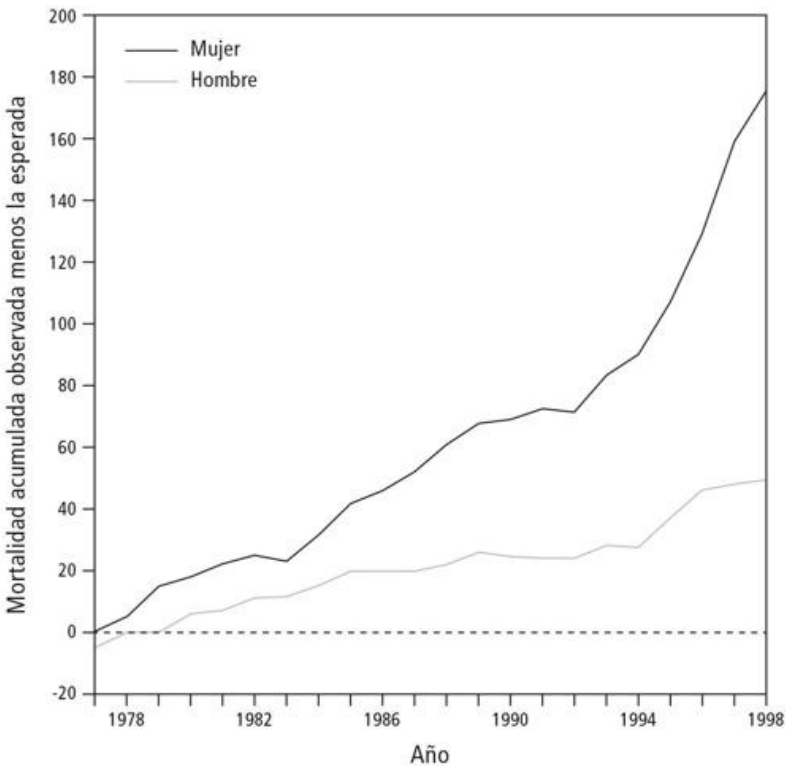


Figura 10.3. Número acumulado de certificados de defunción firmados por Shipman para pacientes de 65 años o más y que murieron en casa o en su clínica. Se ha sustraído, dada la composición de esta lista, el número esperado.

Si esto se hubiese realizado usando las muertes totales para hombres y

mujeres mostradas en la figura 10.3, entonces en 1979, después de solo tres años de supervisión, habría habido un valor P de una cola de 0,004, derivado de la comparación de 40 muertes observadas cuando solo se esperarían 25,3.[141] El resultado podría haber sido declarado «estadísticamente significativo» y Shipman investigado y descubierto.

Pero hay dos razones por las cuales un procedimiento estadístico como ese habría sido sumamente inapropiado como forma de supervisar las tasas de mortalidad de los médicos de cabecera. En primer lugar, a menos que hubiese alguna otra razón para sospechar de Shipman y poner en marcha una investigación centrada en él, tendríamos que haber calculado esos valores P para todos los médicos de cabecera del Reino Unido: alrededor de 25.000 en esa época. Como hemos visto en el ejemplo del salmón muerto, sabemos que si llevamos a cabo suficientes pruebas de significatividad, obtendremos resultados falsos. Con 25.000 médicos de cabecera y un umbral crítico de 0,05, esperaríamos que 1 de cada 20 médicos completamente inocentes —alrededor de 1.300— tuviesen unas cifras de fallecimientos «significativamente altas» cada vez que hiciésemos la prueba, y sería completamente inapropiado investigar a toda esa gente. Y, además, podría perderse el rastro de Shipman en medio de todos estos falsos positivos.

Una alternativa es aplicar el método de Bonferroni, y demandar un valor P de 0,05 / 25.000, o 1 entre 500.000, para el caso más extremo de los médicos de cabecera. Para Shipman, esto habría pasado en 1984, cuando tuvo 105 muertes, frente a las 59,2 que esperaríamos, un exceso de 46.

Pero ni siquiera este podría ser un procedimiento adecuado para todos los médicos de cabecera del país. Un segundo problema se deriva de que estaríamos realizando repetidas pruebas de significatividad, dado que cada año se añaden nuevos datos y se realiza una nueva prueba. Existe una teoría notable pero compleja, conocida con el delicioso término de «ley del logaritmo iterado», que muestra que si llevamos a cabo esas pruebas repetidas, incluso si la hipótesis nula es cierta, es seguro que al final la rechazaríamos sea cual sea el nivel de significatividad que escojamos.

Esto es de lo más preocupante: significa que si continuamos haciendo pruebas a un doctor durante el tiempo suficiente, está garantizado que acabaremos pensando que hemos encontrado evidencia de un exceso de mortalidad, incluso si en realidad sus pacientes no están sometidos a ningún riesgo excesivo. Afortunadamente, hay métodos estadísticos para tratar con este problema de pruebas secuenciales, desarrollados

por vez primera en la Segunda Guerra Mundial por equipos de estadísticos que no estaban implicados en tratamientos médicos, sino en el control de calidad de armamentos y otros materiales de guerra.

Las armas que salían de las cadenas de producción eran objeto de pruebas para comprobar que se adecuaban a un estándar, y todo el proceso era supervisado mediante la acumulación sistemática de las desviaciones totales con respecto al estándar, en gran medida igual que la supervisión del exceso de muertes. Estos científicos se dieron cuenta de que la ley del logaritmo iterado suponía que las repetidas pruebas de significatividad siempre llevarían finalmente a hacer saltar todas las alarmas y a pensar que el proceso industrial estaba fuera de control, aunque en realidad todo estuviese funcionando bien. Estadísticos en Estados Unidos y Reino Unido, trabajando de manera independiente, desarrollaron la que llegó a ser conocida como prueba de razón de probabilidad secuencial (PRPS), que es un estadístico que evalúa la evidencia acumulada sobre desviaciones, y puede en todo momento ser comparado con determinados umbrales: cuando uno de esos umbrales es traspasado, entonces se dispara una alarma y se investiga la línea de producción.[142] Esas técnicas llevaron a unos procesos industriales más eficientes, y fueron más adelante adaptadas para su uso en los llamados ensayos clínicos secuenciales en los que resultados acumulados eran repetidamente analizados para ver si se había alcanzado un umbral que indicase que se había llegado a un tratamiento beneficioso.

Yo formé parte de un equipo que desarrolló una versión de la PRPS que podía aplicarse a los datos de Shipman. Se muestra en la figura 10.4 para hombres y mujeres, asumiendo la hipótesis alternativa de que la tasa de mortalidad de Shipman era el doble que la de sus colegas. La prueba tiene unos umbrales que controlan la probabilidad de error de tipo I (alfa) y tipo II (beta) a los valores especificados de 1 de cada 100, 1 de cada 10.000 y 1 de cada 1.000.000. El error de tipo I es la probabilidad global de que la prueba estadística cruce el umbral en algún momento, dado que Shipman tiene las tasas de mortalidad esperadas, y el error de tipo II es la probabilidad global de que la prueba estadística no cruce el umbral en algún momento, dado que Shipman tiene el doble de la tasa de mortalidad esperada.[143]

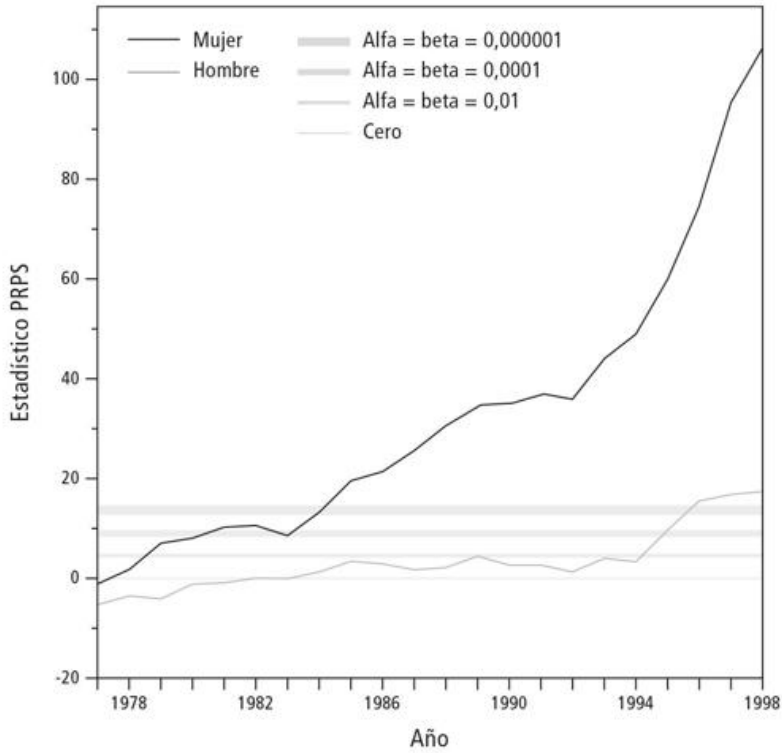


Figura 10.4. Prueba de razón de probabilidad secuencial (PRPS) para la detección de una duplicación de la tasa de mortalidad: pacientes con edad >64 y fallecidos en su casa o clínica. Las líneas rectas indican umbrales que hacen que salte la alarma, que proporcionan las tasas globales de errores de tipo I (alfa) y tipo II (beta) que se muestran —se asumen que son iguales—. Si nos fijamos en la línea para las mujeres, parece claro que Shipman habría cruzado el umbral exterior en 1985.

Dado que hay alrededor de 25.000 médicos de familia, un umbral que suponga un valor P de 0,05 / 25.000, o 1 entre 500.000, podría resultar razonable. Para el caso de las mujeres, Shipman habría cruzado el más estricto umbral alfa = 0,000001, o 1 entre un millón, y si se combinan hombres y mujeres, lo habría hecho en 1984. De manera que, en este caso, la prueba secuencial correcta habría hecho saltar las alarmas al mismo tiempo que la más cándida prueba repetida de significatividad.

Nuestra conclusión para la investigación pública fue que si alguien hubiese estado realizando esta supervisión, y Shipman hubiese sido investigado en 1984 y juzgado, se habrían salvado unas 175 vidas. Solo con aplicar un procedimiento estadístico sencillo de supervisión.

Con el tiempo, se llevó a cabo una prueba con un sistema de supervisión a los médicos de familia, ¡que inmediatamente identificó a un médico de familia con tasas de mortalidad aún mayores que las de Shipman! La investigación reveló que este doctor ejercía en una ciudad de la costa sur con un gran número de residencias de la tercera edad con muchas personas mayores, y se esmeraba en ayudar a sus pacientes a que no muriesen en el hospital. Habría sido completamente inapropiado que este médico de familia recibiese ningún tipo de publicidad por la tasa aparentemente alta de certificados de defunción que había firmado. La lección que se puede extraer de todo esto es que aunque los sistemas estadísticos pueden detectar resultados atípicos, no pueden ofrecer las razones que explican por qué se dan esos resultados, por lo que se requiere aplicarlos cuidadosamente para evitar acusaciones falsas. Una razón más para tener cuidado con los algoritmos.

¿Qué puede salir mal con los valores P?

Ronald Fisher desarrolló la idea del valor P como una medida de la compatibilidad de los datos con alguna hipótesis previa. Por ello, si se calcula un valor P y se encuentra que es pequeño, esto significa que si la hipótesis fuese cierta, es poco probable que el estadístico de resumen sea tan extremo, de manera que o bien algo sorprendente ha ocurrido, o bien la hipótesis original es defectuosa. La lógica podría parecer algo enrevesada, pero ya hemos visto lo útil que puede resultar esa idea básica. Teniendo esto en cuenta, ¿qué podría salir mal?

Pues resulta que muchas cosas. Fisher se planteó el tipo de situación que hemos visto en los ejemplos anteriores de este capítulo, con un conjunto único de datos, una única medida de resumen de los resultados y una prueba única de compatibilidad. Pero en las últimas décadas los valores P se han convertido en moneda corriente en toda investigación; la literatura científica los utiliza masivamente —un estudio llegó a identificar alrededor de treinta mil estadísticos-t con sus respectivos valores P en solo tres años de artículos en dieciocho revistas de psicología y neurociencia—.[144]

Pongamos como ejemplo de lo que puede ocurrir un conjunto de, digamos, 1.000 estudios, cada uno diseñado con una significatividad del 5 % (α) y un 80 % de potencia ($1 - \beta$), una potencia considerablemente superior que la mayor parte de los estudios. En el mundo real de la investigación, aunque los experimentos se llevan a cabo con la esperanza de hacer un descubrimiento, se reconoce que la mayor parte de las hipótesis nulas son (al menos aproximadamente)

ciertas. Supongamos que solo un 10 % de las hipótesis nulas analizadas son falsas. Esta cifra es incluso probablemente muy alta para nuevos productos farmacéuticos, que tienen una tasa de éxitos notoriamente baja. Al igual que los ejemplos mostrados en el capítulo 8, la figura 10.5 muestra las frecuencias para lo que esperamos que ocurra en estos 1.000 estudios.

Esto indica que esperamos encontrar 125 «descubrimientos», pero de esos, 45 son falsos positivos: en otras palabras, el 36 %, es decir, alrededor de un tercio de las hipótesis nulas rechazadas (los «descubrimientos») son afirmaciones incorrectas. Esta visión tan deprimente incluso empeora cuando vamos a la literatura científica, dado que las revistas están sesgadas a favor de publicar resultados positivos. Un análisis similar de estudios científicos llevó a John Ioannidis, profesor de Medicina y Estadística de Stanford, a proclamar célebremente que «la mayoría de los descubrimientos científicos publicados son falsos».[145] Volveremos a las razones de esta conclusión devastadora en el capítulo 12.

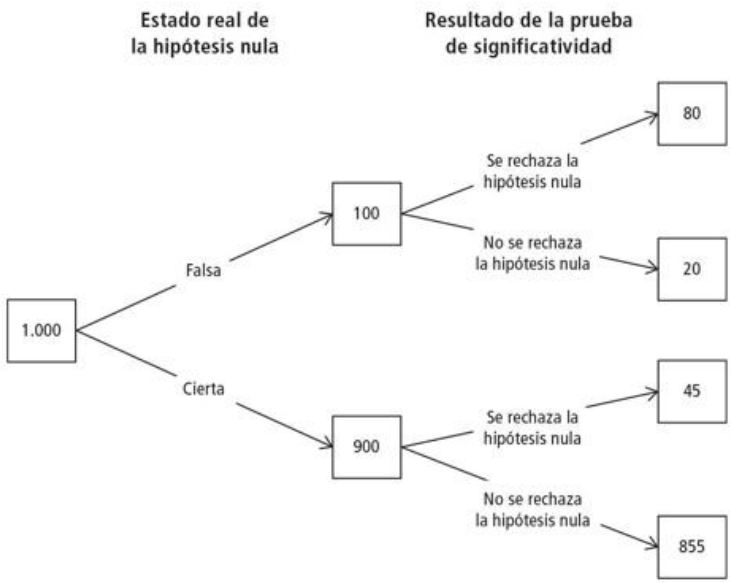


Figura 10.5. Frecuencias esperadas de los resultados de 1.000 pruebas de hipótesis llevadas a cabo con significatividad del 5 % (error tipo I, α) y 80 % de potencia (1 – error de tipo II, 1 – β). Solo el 10 % (100) de las hipótesis nulas son falsas, y detectamos correctamente el 80 % de ellas (80). De las 900 hipótesis nulas que son ciertas, rechazamos incorrectamente 45 (5 %). En general, de los 125 «descubrimientos», el 36 % (45) son falsos descubrimientos.

Dado que todos estos falsos descubrimientos se basaban en un valor P que identificaba un resultado «significativo», se ha responsabilizado cada vez más a los valores P por haber provocado una riada de conclusiones científicas incorrectas. En 2015, una reputada revista de psicología llegó a anunciar que prohibiría el uso del TSHN (test de significatividad de la hipótesis nula). Finalmente, en 2016, la Asociación Americana de Estadística (ASA)[146] logró reunir a un grupo de estadísticos que acordaron seis principios sobre los valores P. [147]

El primero de estos principios simplemente señala lo que los valores P pueden hacer:

1. Los valores P pueden indicar cómo son de incompatibles los datos con un determinado modelo estadístico.

Como hemos visto repetidas veces, los valores P consiguen este efecto a base de medir hasta qué punto los datos son sorprendentes, dada una hipótesis nula de que algo no existe. Por ejemplo, cuando nos preguntamos si los datos son incompatibles con un medicamento que no funciona. La lógica puede ser algo compleja, pero es útil.

El segundo principio intenta remediar errores en su interpretación:

2. Los valores P no miden la probabilidad de que la hipótesis estudiada sea cierta, o la probabilidad de que los datos sean resultado exclusivamente del azar.

En el capítulo 8 vimos que había que tener mucho cuidado al distinguir afirmaciones correctas referidas a una probabilidad condicional del tipo «solo un 10 % de las mujeres sin cáncer de mama obtendrá una mamografía positiva» de afirmaciones incorrectas del tipo «solo un 10 % de las mujeres con una mamografía positiva no tienen cáncer de mama». Este es el error conocido como «falacia del fiscal», y vimos que hay formas de evitarlo por completo, centrándonos en qué podríamos esperar que les pase a 1.000 mujeres que hayan sido sometidas a la mamografía.

Se pueden producir problemas similares en el caso de los valores P, que miden la probabilidad de que se den unos datos muy extremos, si la hipótesis nula es cierta, y no la probabilidad de que la hipótesis nula sea cierta, dado que unos datos tan extremos han ocurrido. Esta es una diferencia sutil pero esencial.

Cuando los equipos del CERN declararon haber obtenido un resultado «cinco sigma» para el bosón de Higgs, lo que correspondía a un valor

P de alrededor de 1 en 3,5 millones, la BBC informó de manera correcta que esto significaba «una probabilidad de alrededor de 1 entre 3,5 millones de que la señal que han visto apareciese aunque no haya partícula de Higgs». Pero casi todos los demás medios interpretaron mal el significado de este valor P. Por ejemplo, Forbes Magazine informó: «La probabilidad es menor a 1 entre 1 millón de que no sea el bosón de Higgs», un claro ejemplo de falacia del fiscal. La interpretación del Independent fue la más habitual; informó de que «hay una probabilidad de menos de 1 entre 1 millón de que sus resultados sean pura chiripa estadística». Esto no es tan rotundamente erróneo como la afirmación de Forbes, pero sigue suponiendo asignar esa pequeña probabilidad al hecho de que «sus resultados sean pura chiripa estadística», que es lógicamente lo mismo que decir que es la probabilidad de que la hipótesis nula sea comprobada. Por eso mismo la ASA intenta enfatizar que el valor P no es «la probabilidad de que los datos sean producidos únicamente por puro azar aleatorio».

El tercer principio de la ASA intenta contrarrestar la obsesión con la significatividad estadística:

3. Las conclusiones científicas y las decisiones económicas o políticas no deberían basarse en si un valor P cruza un determinado umbral.

Cuando Ronald Fisher comenzó a publicar tablas que mostraban valores de estadísticos con «$P < 0,05$» o «$P < 0,01$», presumiblemente no tenía mucha idea de hasta qué punto esos umbrales bastante arbitrarios llegarían a dominar las publicaciones científicas, en las que todos los resultados tienden al final a separarse en «significativos» o «no significativos». A partir de aquí solo hay un pequeño paso hasta considerar los resultados «significativos» como descubrimientos demostrados, lo que produce un precedente simplificado y peligroso de cómo ir de los datos directamente a las conclusiones sin pararse a pensar en lo que se está haciendo.

Una consecuencia de lo más calamitosa de emplear esta dicotomía simple es la tergiversación de lo que significa «no significativo». Un valor P no significativo sugiere que los datos son compatibles con la hipótesis nula, pero esto no significa que la hipótesis nula sea cierta. Después de todo, que no haya evidencia directa de que un criminal haya estado en la escena del crimen no significa que sea inocente. Pero este error es sorprendentemente común.

Considérese la gran disputa científica sobre si una pequeña cantidad de alcohol, por ejemplo una copa al día, es buena para usted. Un estudio afirmó que solo las mujeres mayores podrían beneficiarse de un consumo moderado de alcohol, pero un análisis más detallado mostró que había otros grupos que también mostraban beneficios, pero cuyos resultados no eran estadísticamente significativos, dado que los intervalos de confianza alrededor de los beneficios estimados para estos grupos eran ciertamente muy amplios. Aunque los intervalos de confianza incluían el 0 y, por tanto, los efectos no eran estadísticamente significativos, los datos eran completamente compatibles con la sugerida reducción de entre un 10 % y un 20 % en el riesgo de fallecimiento. Pero The Times anunció a los cuatro vientos que «el alcohol no tiene beneficios para la salud después de todo». [148]

En resumen, es muy engañoso interpretar que «no es significativamente distinto de 0» signifique que el efecto real es 0, particularmente en estudios pequeños con poca potencia y amplios intervalos de confianza.

El cuarto principio de la ASA suena bastante inocuo:

4. Una inferencia apropiada requiere informar de todo con transparencia.

La necesidad más obvia es informar claramente de cuántas pruebas se han hecho, de tal manera que si se enfatiza el resultado más significativo, podemos aplicar algún tipo de ajuste, como el de Bonferroni. Solo sabiendo cuál es el plan del estudio, y qué es lo que se ha hecho realmente, pueden evitarse problemas con los valores P.

Ha planificado usted su estudio, reunido sus datos, hecho sus análisis y obtenido un resultado «significativo». Sin duda es este un descubrimiento importante, ¿no? El quinto principio de la ASA recomienda no ser demasiado arrogante:

5. Un valor P, o significatividad estadística, no mide el tamaño de un

efecto o la importancia de un resultado.

Nuestro siguiente ejemplo muestra que, particularmente si tenemos grandes muestras, podríamos estar razonablemente confiados de que existe una asociación, pero aun así no muy impresionados por la importancia de esta.

¿Por qué ir a la universidad incrementa el riesgo de desarrollar un tumor cerebral?

Vimos este titular en el capítulo 4. Después de ajustar por estatus matrimonial e ingresos en un análisis de regresión, los investigadores suecos encontraron un incremento relativo del riesgo del 19 % entre el nivel educativo más bajo (solo primaria) y el más alto (grado universitario), con un intervalo de confianza del 95 % que iba de 7 % a 33 %; es notable que el artículo no informaba de los valores P, pero dado que el intervalo de confianza del 95 % para el riesgo relativo excluye el 1, podemos concluir que $P < 0{,}05$.

A estas alturas, el lector ya debería estar listo para enfrentarse a una lista de potenciales inquietudes referidas a las conclusiones del estudio en cuestión, pero los autores ya las habían previsto. Junto con los resultados de su estudio, reconocieron que:

→ No se podía hacer una interpretación causal.

→ No se había controlado por variables de confusión potenciales asociadas al estilo de vida, como el consumo de alcohol.

→ Es probable que las personas con alto estatus socioeconómico tengan una mayor tendencia a buscar atención médica, por lo que podría haber un sesgo.

Pero no se mencionó una cuestión importante: el pequeño tamaño de la aparente asociación. Un incremento del 19 % entre los niveles educativos bajo y alto es mucho menor de lo que se encuentra en muchos cánceres. El artículo señaló que se habían diagnosticado 3.715 tumores en alrededor de 2.000.000 hombres a lo largo de 18 años (aproximadamente 1 de 600), de manera que, siguiendo el procedimiento presentado en el capítulo 1 para traducir el riesgo relativo en cambios en el riesgo absoluto, podemos calcular que:

→ De alrededor de 3.000 con el nivel educativo más bajo, podríamos esperar que se diagnosticasen unos 5 tumores (un riesgo base de 1 entre 600).

→ De 3.000 hombres con el nivel educativo más alto, esperaríamos 6 (un incremento relativo del 6 %).

Esto nos da una impresión algo diferente acerca de los resultados, y de hecho, es una impresión bastante alentadora. Un incremento tan pequeño en el riesgo de un cáncer poco habitual solo puede ser estadísticamente significativo cuando se estudia un número enorme de personas: en este caso, alrededor de un millón de hombres.

Las principales lecciones que se pueden extraer de este estudio científico podrían ser, por tanto, (a) que los «macrodatos» pueden fácilmente llevar a resultados que son estadísticamente significativos pero que no tienen significatividad práctica, y (b) que no debe usted preocuparse de que estudiar para sacarse la carrera le vaya a provocar un tumor cerebral.

El último principio de la ASA es bastante más sutil:

6. En sí mismo, el valor P no proporciona una buena medida de la evidencia referida a un modelo o hipótesis. Por ejemplo, un valor P cercano a 0,05, tomado aisladamente, solo ofrece una evidencia débil contra la hipótesis nula.

Esta afirmación, en parte basada en el razonamiento «bayesiano» que expondremos en el capítulo siguiente, ha llevado a un prominente grupo de estadísticos a argumentar que el umbral estándar para el «descubrimiento» de un nuevo efecto debería cambiarse a $P < 0{,}005$. [149]

¿Qué efecto puede tener esto? Cambiar el criterio de «significatividad» de 0,05 (1 de cada 20) a 0,005 (1 de cada 200) en la figura 10.5 supondría que en lugar de tener 45 «descubrimientos» falsos positivos, tendríamos solo 4,5. Esto reduciría el número total de descubrimientos a 84,5, y de estos solo 4,5 (5 %) serían descubrimientos falsos. Lo que sería una mejora considerable frente al 36 %.

La idea original de Fisher para comprobar hipótesis ha sido muy beneficiosa para la práctica de la estadística y la prevención de afirmaciones científicas no justificadas. Pero los estadísticos se han quejado frecuentemente de la facilidad con que algunos investigadores pasan sin mayor problema de unos valores P obtenidos a partir de estudios malamente diseñados a inferencias generalizables: un tipo de alquimia que convierte lo incierto en seguro, aplicando mecánicamente pruebas estadísticas para dividir todos los estudios en «significativos» y «no significativos». Veremos algunas de las malas

consecuencias derivadas de ello en el capítulo 12, pero primero nos referiremos a un enfoque alternativo a la inferencia estadística que rechaza por completo la idea misma de los test de significatividad de la hipótesis nula.

Para ello tendremos que hacer otro ejercicio mental de esos a los que nos tiene acostumbrados la ciencia estadística, y olvidar todo lo que podamos haber aprendido de este y los anteriores capítulos.

Resumen

→ Las pruebas de las hipótesis nulas —asunciones por defecto sobre modelos estadísticos— son una parte esencial de la práctica estadística.

→ Un valor P es una medida de la incompatibilidad entre los datos observados y una hipótesis nula: formalmente es la probabilidad de observar un resultado extremo, si la hipótesis nula fuese cierta.

→ Tradicionalmente, se han establecido umbrales para los valores P de 0,05 y 0,01 para declarar que algo es «estadísticamente significativo».

→ Estos umbrales necesitan ser ajustados si se llevan a cabo múltiples pruebas, por ejemplo sobre diferentes subconjuntos de los datos o múltiples medidas de los resultados.

→ Hay una correspondencia precisa entre los intervalos de confianza y los valores P: si, digamos, el intervalo del 95 % excluye el 0, podemos rechazar la hipótesis nula de 0 para $P < 0{,}05$.

→ La teoría Neyman-Pearson especifica una hipótesis alternativa, y fija las tasas de errores de tipo I y tipo II para los dos tipos posibles de errores en una prueba de hipótesis.

→ Se han desarrollado formas separadas de pruebas de hipótesis para las pruebas secuenciales.

→ Los valores P son a menudo malinterpretados; en particular, no transmiten la probabilidad de que la hipótesis nula sea cierta, ni tampoco un resultado no significativo implica que la hipótesis nula sea cierta.

[129] J. Arbuthnot, «An Argument for Divine Providence…», Philosophical Transactions 27 (1710), pp. 186-190.

[130] R. A. Fisher, The Design of Experiments, Oliver and Boyd, 1935, p. 19.

[131] Quizá una pregunta más obvia sería la relación entre cruzar los brazos y el ser zurdo o diestro, pero había muy pocos zurdos para investigarla.

[132] Podríamos escoger otro estadístico que resuma las asociaciones, como la razón de probabilidad, pero obtendríamos esencialmente el mismo resultado.

[133] Hay $54 \times 53 \times 52 \ldots \times 2 \times 1$ permutaciones, lo que se denomina «54 factorial», expresado como 54!. Esto es aproximadamente 2 con 71 ceros detrás. Nótese que el número de formas posibles en que un taco de 52 cartas puede ser repartido es 52!, por lo que incluso si repartimos un millón de millones de manos por segundo, el número de años que nos llevaría realizar todas las permutaciones tiene 48 ceros detrás, mientras que la edad del universo es de solo 14.000.000.000 años. Es por ello que podemos estar bastante seguros de que a lo largo de toda la historia de los juegos de cartas no hay dos tacos que hayan sido mezclados en el mismo orden exactamente.

[134] «Chi» (pronunciado «ji») es la letra griega χ.

[135] El truco es calcular, para cada uno de los 412 individuos en el conjunto de validación, la diferencia entre los errores de predicción al cuadrado para los dos algoritmos; este conjunto de 412 diferencias tiene una media de $-0{,}0027$ y una desviación típica de 0,1028. El error típico de la estimación de la diferencia «auténtica», por tanto, es $0{,}1028/\sqrt{412} = 0{,}0050$, y el estadístico t el estimador/error típico $= -0{,}0027/0{,}0050 = -0{,}54$. Esto se conoce como una prueba t emparejada dado que se basa en un conjunto de diferencias entre pares de números.

[136] La probabilidad exacta de que al menos uno de los ensayos fuese significativo es 1 − (la probabilidad de que ambos sean no significativos) $= 1 - 0{,}95 \times 0{,}95 =$

$= 0{,}0975$. que se redondea a 0,10.

[137] El estudio del pez muerto está descrito en este póster: http://prefrontal.org/files/posters/Bennett-Salmon-2009.jpg.

[138] El anuncio del CERN del bosón de Higgs está en: http://cms.web.cern.ch/news/observation-new-particle-mass-125-gev.

[139] La teoría original de Neyman y Pearson incluía la idea de «aceptar» una hipótesis nula, pero esta parte de su teoría es ahora ignorada.

[140] D. Spiegelhalter, O. Grigg, R. Kinsman y T. Treasure, «Risk-Adjusted Sequential Probability Ratio Tests: Applications to Bristol, Shipman and Adult Cardiac Surgery», International Journal for Quality in Health Care 15 (2003), pp. 7-13.

[141] El valor P es de una sola cola dado que solo estamos interesados en detectar una mayor mortalidad, y no un descenso. El valor P es, por tanto, la probabilidad de que una variable aleatoria Poisson con media 22,5 sea al menos 40, que en el software convencional sería 0,004.

[142] Los estadísticos fueron dirigidos por Abraham Wald en Estados Unidos y George Barnard en el Reino Unido. Barnard era un hombre encantador, un matemático puro (y comunista) antes de la guerra, cuando como muchos otros adaptó sus conocimientos al trabajo estadístico en favor del esfuerzo bélico. Más adelante desarrollaría el estándar británico oficial de los condones (BS 3704).

[143] La prueba estadística tiene la siguiente forma sencilla: PRPS = 0,69 × muertes observadas acumuladas − muertes esperadas acumuladas. Los umbrales vienen dados por log ((1 − β) / α).

[144] D. Szucs y J. P. A. Ionnidis, «Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature», PLOS Biology 15:3 (2 de marzo de 2017), e2000797.

[145] J. P. A. Ioannidis, «Why Most Published Research Findings Are False», PLOS Medicine 2:8 (agosto de 2005), e124.

[146] Por sus siglas en inglés, American Statistical Association. (N. del T.).

[147] Fue este un logro notable, dado que el nombre colectivo aplicado a los estadísticos es «varianza».

[148] C. S. Knott et al., «All Cause Mortality and the Case for Age Specific Alcohol Consumption Guidelines: Pooled Analyses of up to 10 Population Based Cohorts», British Medical Journal 350 (10 de febrero de 2015), h384. Y debajo del titular: «Alcohol Has No Health Benefits After All», The Times, 11 de febrero de 2015.

[149] D. J. Benjamin et al., «Redefine Statistical Significance», Nature

Human Behavior 2 (2018), pp. 6-10.

11

Aprender de la experiencia

a la manera bayesiana

«No estoy seguro del todo

de que la “confianza” no sea

un “abuso de confianza”».

ARTHUR BOWLEY, 1934

Me veo obligado a admitir algo en nombre de la comunidad estadística. La base formal del aprendizaje a partir de datos es un poco desastrosa. Aunque ha habido numerosos intentos de producir una única teoría unificada de la inferencia estadística, ninguno de ellos se ha aceptado por completo. No es por casualidad que a los matemáticos no les guste enseñar Estadística.

Ya hemos visto las ideas antagónicas de Fisher y Neyman-Pearson, y ya es hora de analizar un tercer acercamiento a la inferencia, el bayesiano. Este enfoque solo ha adquirido prominencia en los últimos cincuenta años, pero sus principios básicos se remontan algo más, al reverendo Thomas Bayes, un ministro inconformista de Tunbridge Wells convertido en teórico de la probabilidad y filósofo, que murió en 1761.[150]

Las buenas noticias son que el enfoque bayesiano abre nuevas posibilidades para sacar el máximo provecho de los datos más complejos. Las malas son que esto significa olvidarse de casi todo lo que usted podría haber aprendido en este libro y en cualquier otro sitio sobre estimación, intervalos de confianza, valores P, pruebas de hipótesis y demás.

¿Qué es el enfoque bayesiano?

La primera gran contribución de Thomas Bayes fue el uso de la probabilidad como una expresión de nuestra falta de conocimientos sobre el mundo o, lo que es lo mismo, de nuestra ignorancia acerca de qué está pasando. Mostró que la probabilidad puede emplearse no solo para eventos futuros sujetos al azar —incertidumbre aleatoria, por usar el término introducido en el capítulo 8—, sino también para eventos que son ciertos, y podrían ser bien conocidos por algunas

personas, pero desconocidos para nosotros: la llamada incertidumbre epistémica.

Si lo piensa, estamos rodeados de incertidumbre epistémica sobre cosas que son invariables pero nos son desconocidas. Los jugadores apuestan sobre la siguiente carta que repartir, compramos billetes de lotería, discutimos sobre el posible sexo de un bebé, especulamos sobre historias policiacas, discutimos sobre el número de tigres que quedan en la selva y nos dan estimaciones del posible número de migrantes (o de desempleados). Todos estos son hechos o cantidades que existen ahí fuera, pero no sabemos cuáles son. Por decirlo una vez más, desde una perspectiva bayesiana, está bien usar las probabilidades para representar nuestra ignorancia personal sobre estos hechos y números. Podríamos incluso pensar en asignar probabilidades a teorías científicas alternativas, pero esto es más discutible.

Estas probabilidades, por supuesto, dependerán de nuestro conocimiento actual: recordemos del capítulo 8 cómo nuestra probabilidad acerca de si al lanzar una moneda ha salido cara o cruz depende de si la hemos mirado o no. De manera que estas probabilidades bayesianas son necesariamente subjetivas: dependen de nuestra relación con el mundo exterior, y no son propiedades del mundo en sí mismo. Deberían cambiar a medida que recibimos nueva información.

Lo que nos lleva a la segunda contribución clave de Bayes: la posibilidad de revisar continuamente nuestras probabilidades actuales a la luz de nueva evidencia. Esto se conoce como teorema de Bayes, y proporciona esencialmente un mecanismo formal para aprender de la experiencia, lo cual es un logro extraordinario para un clérigo desconocido de una pequeña ciudad balneario inglesa. El legado de Bayes es la idea fundamental de que los datos no hablan por sí mismos: nuestro conocimiento externo, e incluso nuestro juicio, juegan un papel central. Esto puede parecer incompatible con el proceso científico, pero por supuesto el conocimiento y la comprensión del contexto siempre han sido elementos que tener en cuenta en el aprendizaje a partir de datos. La diferencia es que en el enfoque bayesiano esto es algo que se maneja de manera formal y matemática.

Las implicaciones del trabajo de Bayes han sido fuertemente cuestionadas. Muchos estadísticos y filósofos han puesto objeciones a la idea de que los juicios subjetivos tengan un papel en la ciencia estadística. Por ello, es justo que deje clara cuál es mi posición

personal: me formé en la escuela «subjetivista» bayesiana de razonamiento estadístico al comienzo de mi carrera,[151] y, para mí, sigue siendo el enfoque más satisfactorio.

Tiene usted tres monedas en su bolsillo: una tiene dos caras, otra tiene cara y cruz y la tercera tiene dos cruces. Coge una moneda al azar, la lanza y sale cara. ¿Cuál sería la probabilidad de que el otro lado de la moneda sea también cara?

Este es un problema clásico de incertidumbre epistémica: no hay aleatoriedad en la moneda una vez que se ha lanzado, y cualquier probabilidad es simplemente una expresión de su actual ignorancia personal sobre el otro lado de la moneda.

Mucha gente se apresuraría a afirmar que la respuesta es ½, dado que la moneda debe ser o la de cara y cruz o la de dos caras, y cada una tiene la misma probabilidad de ser escogida. Hay muchas formas de evaluar si esto es correcto, pero la más sencilla es usar la idea de frecuencias esperadas demostrada en el capítulo 8.

La figura 11.1 muestra qué esperaríamos ver si llevamos a cabo este ejercicio seis veces. De media, cada moneda sería escogida dos veces, y cuando la lanzamos, saldría cada lado de cada moneda. En tres de los casos saldría cara, y en dos de estos tres sería con una moneda de dos caras. De manera que la probabilidad de que la moneda escogida sea de dos caras en lugar de tener cara y cruz debería ser de ⅔ y no de ½. Esencialmente, que salga cara hace más probable que se haya escogido una monera con dos caras, dado que esta moneda proporciona dos oportunidades para que salga cara, mientras que la moneda con cara y cruz solo proporciona una.

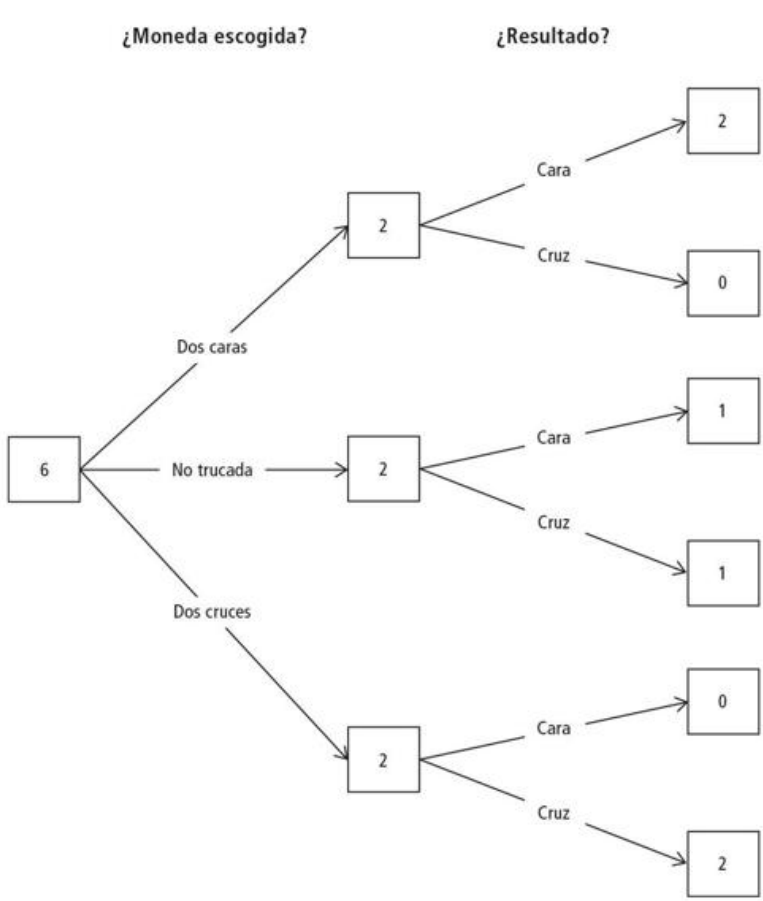


Figura 11.1. Árbol de frecuencia esperada para el problema de las tres monedas, que muestra lo que esperaríamos que ocurriese en las seis repeticiones.

Si este resultado le parece contraintuitivo, el siguiente ejemplo le va a parecer aún más sorprendente.

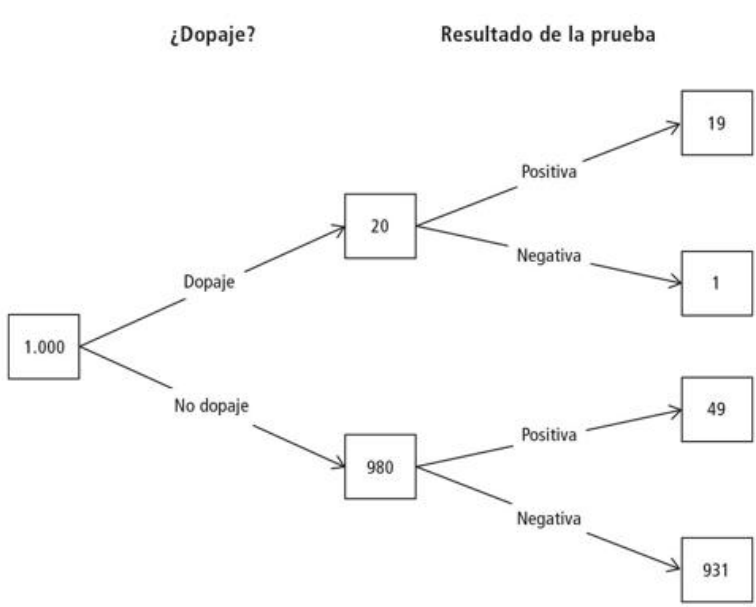


Figura 11.2. Árbol de frecuencia esperada para el dopaje en el deporte, que muestra lo que esperamos que les ocurra a 1.000 atletas cuando 1 de cada 50 se dopa, y la prueba de dopaje es un «95 % fiable».

Supongamos que se afirma que una prueba de dopaje para el deporte es un «95 % fiable», lo que significa que un 95 % de deportistas dopados y un 95 % de no dopados serán clasificados correctamente. Asumamos que en cada momento 1 de cada 50 atletas se dopa. Si un atleta da positivo, ¿cuál es la probabilidad de que realmente se esté dopando?

Este tipo de problemas potencialmente complejos es mejor tratarlos mediante frecuencias esperadas, al igual que hicimos con el análisis de las pruebas para cáncer de mama en el capítulo 8, y con las afirmaciones del capítulo 10 de que una alta proporción de la literatura científica publicada es errónea.

El árbol de la figura 11.2 comienza con 1.000 atletas, de los cuales 20 se dopan y 980 no. Todos excepto uno de ellos son detectados (95 % de 20 = 19), pero 49 que no se dopan dan también positivo en las pruebas (95 % de 980 = 931). Esperaríamos por tanto un total de 19 + 49 = 68 pruebas positivas, de las cuales solo 19 son casos reales de dopaje —el otro 72 % de pruebas positivas son acusaciones falsas—. Incluso aunque las pruebas de dopaje sean realmente «fiables al 95 %», la mayoría de los que dan positivo son, de hecho, inocentes; no se requiere mucha imaginación para ver los problemas que esta aparente

paradoja causaría en la vida real, en la que los atletas son rutinariamente condenados porque han dado positivo en una prueba de dopaje.

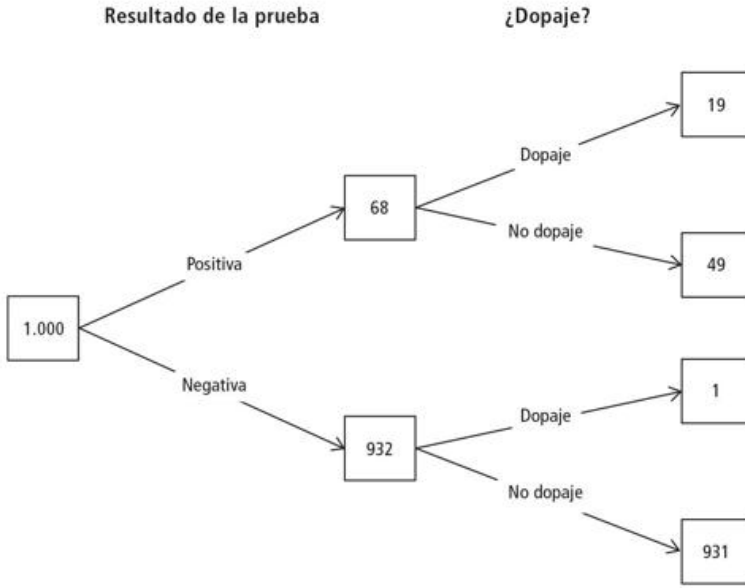


Figura 11.3. Árbol de frecuencia esperada inversa para el dopaje en el deporte, restructurado de manera que primero vienen los resultados de la prueba, seguidos de la revelación de la actividad real del atleta.

Una forma de ver este proceso es que estamos «invirtiendo el orden» del árbol para poner primero la prueba y después la revelación de la verdad. Esto se muestra explícitamente en la figura 11.3.

Este «árbol invertido» llega a exactamente los mismos resultados finales, pero respeta el orden temporal en el que nos enteramos de lo que pasa (primero la prueba y luego la verdad sobre el dopaje), en lugar de lo que sería la secuencia temporal de la causalidad subyacente (primero el dopaje y luego la prueba). Esta «inversión» es exactamente lo que hace el teorema de Bayes; de hecho el pensamiento bayesiano era conocido como «probabilidad inversa» hasta la década de 1950.

El ejemplo del dopaje en el deporte muestra lo fácil que es confundir la probabilidad de haberse dopado, dada una prueba positiva (28 %), con la probabilidad de dar positivo, dado que se ha dopado (95 %). Ya hemos visto otros contextos en los que la probabilidad de «A dado B» se confunde con la probabilidad de «B dado A»:

→ La interpretación errónea de los valores P, según la cual la probabilidad de la evidencia dada la hipótesis nula se confunde con la probabilidad de la hipótesis nula dada la evidencia.

→ La falacia del fiscal en los casos judiciales, por la cual la probabilidad de la evidencia dado que el acusado es inocente se confunde con la probabilidad de que el acusado sea inocente dada la evidencia.

Un observador razonable podría pensar que el pensamiento formal bayesiano llevaría claridad y rigor al modo de manejar la evidencia en los casos legales. Y se sorprendería al enterarse de que el teorema de Bayes está esencialmente prohibido en los tribunales británicos. Antes de revelar cuáles son los argumentos detrás de esta prohibición, debemos echar primero un vistazo a la magnitud estadística que sí está permitida en los tribunales: la razón de verosimilitud.

Razones de probabilidad y de verosimilitud

El ejemplo del dopaje nos permite dar los pasos lógicos necesarios para llegar al número que realmente nos interesa cuando se toman decisiones: de todas las personas que dan positivo, la proporción de los que realmente se dopan, que resulta ser 19/68. El árbol de frecuencia esperada muestra que esto depende de tres números cruciales: la proporción de atletas que se dopan (1/50, o 20/1.000 en el árbol), la proporción de atletas dopados que dan positivo (95 %, o 19/20 en el árbol) y la proporción de atletas que no se dopan que erróneamente dan positivo (5 %, o 49/980 en el árbol).

El análisis se vuelve (bastante) intuitivo usando un árbol de frecuencia esperada, aunque el teorema de Bayes también puede ser expresado mediante una fórmula conveniente usando probabilidades. Pero primero debemos volver a la idea de la razón de probabilidad introducida en el capítulo 1, una idea que no necesita ser explicada a los jugadores experimentados, al menos si son británicos. La razón de probabilidad de un evento es la probabilidad de que ocurra, dividida por la probabilidad de que no ocurra. De manera que la razón de probabilidad de lanzar una moneda y que salga cara es 1, el resultado de dividir ½ (la probabilidad de que salga cara) entre ½ (la probabilidad de que salga cruz).[152] La razón de probabilidad de lanzar un dado y sacar un seis es 1/6 dividido entre 5/6, lo que nos da 1/5, conocido popularmente como «favorables 1 a 5», o «desfavorables 5 a 1», si usted usa el método británico para expresar razones de probabilidad en apuestas.

A continuación necesitamos introducir la idea de la razón de verosimilitud, un concepto que se ha vuelto fundamental para comunicar la solidez de la evidencia forense en casos penales. Cada vez es más frecuente que los jueces y los abogados reciban la formación necesaria para comprender las razones de verosimilitud, que esencialmente comparan el apoyo relativo proporcionado por una determinada evidencia para dos hipótesis rivales, que llamaremos A y B, pero que normalmente representan culpabilidad o inocencia. Técnicamente, la razón de verosimilitud es la probabilidad de que la evidencia asuma la hipótesis A, dividida por la probabilidad de que la evidencia asuma la hipótesis B.

Veamos cómo funciona esto en el caso del dopaje, en el que la «evidencia» forense es el resultado positivo de la prueba, la hipótesis A es que el atleta es culpable de dopaje, y la hipótesis B, que es inocente. Asumimos que el 95 % de los que se dopan dan positivo en la prueba, por lo que la probabilidad de la evidencia dada la hipótesis A es 0,95. Sabemos que un 5 % de los que no se dopan dan positivo, por lo que la probabilidad de la evidencia dada la hipótesis B es 0,05. De manera que la razón de probabilidad es 0,95 / 0,05 = 19: es decir, es 19 veces más probable que la prueba dé positivo si el atleta es culpable que si es inocente. Esto en principio podría parecer una evidencia muy clara, pero más adelante nos encontraremos con razones de verosimilitud de millones y de miles de millones.

Todo lo que hemos visto se resume en el teorema de Bayes, que simplemente dice que:

la razón de probabilidad inicial de una hipótesis × la razón de verosimilitud = la razón de probabilidad final de la hipótesis

Para el ejemplo del dopaje, la razón de probabilidad inicial para la hipótesis «el atleta se dopa» es 1/49 y la razón de verosimilitud es 19, por lo que el teorema de Bayes nos diría que la razón de probabilidad final viene dada por:

$1/49 \times 19 = 19/49$

Esta razón de probabilidad de 19/49 puede ser transformada en una probabilidad de 19/(19 + 49) = 19/68 = 28 %. De manera que esta probabilidad, que se obtiene a partir del árbol de frecuencia esperada de manera bastante sencilla, puede también derivarse de la ecuación general del teorema de Bayes.

Usando un lenguaje más técnico, la razón de probabilidad inicial se

conoce como razón de probabilidad «previa», y la final, como razón de probabilidad «posterior». Esta fórmula puede aplicarse repetidamente, pasando la razón de probabilidad posterior a ser la previa cuando se introduce nueva evidencia independiente. La combinación de toda la evidencia es equivalente a multiplicar juntas las razones de verosimilitud para formar una razón de verosimilitud compuesta.

El teorema de Bayes parece engañosamente básico, pero encierra una forma inmensamente poderosa de aprender de los datos.

Razones de verosimilitud y ciencia forense

El sábado 25 de agosto de 2012, los arqueólogos comenzaron la búsqueda de los restos de Ricardo III excavando en un aparcamiento en Leicester. A las pocas horas encontraron un primer esqueleto.

¿Cuál era la probabilidad de que fuese Ricardo III?

En la leyenda popular, promovida por el apologista de la era Tudor William Shakespeare, Ricardo III (el último rey de la casa de York) era un malvado jorobado. Aunque esta visión ha sido muy discutida, sí está demostrado por los registros históricos que murió en la batalla del campo de Bosworth el 22 de agosto de 1485, a los 32 años de edad, y que su muerte puso fin a la guerra de las Rosas. Se dice que su cuerpo fue mutilado y llevado al priorato de Greyfriars en Leicester, que fue más tarde demolido y finalmente cubierto por un aparcamiento.

Teniendo en cuenta simplemente esa información, podríamos asumir que esos huesos serían los restos de Ricardo III si todo lo siguiente fuese cierto:

→ Realmente había sido enterrado en Greyfriars.

→ Su cadáver no había sido desenterrado y trasladado o esparcido en otro lado en los quinientos veintisiete años transcurridos.

→ El primer esqueleto encontrado resultó ser él.

Supongamos que realizamos asunciones bastante pesimistas, y asumamos que solo hay un 50 % de probabilidades de que las historias de su enterramiento sean ciertas, y un 50 % de probabilidades de que su esqueleto esté todavía donde fue originalmente enterrado en Greyfriars. E imaginemos que otros 100 cadáveres fueron también enterrados en ese mismo lugar (los arqueólogos tenían una buena idea de donde cavar, dado que se decía que Ricardo había sido enterrado en el coro del convento). Si esto es

así, la probabilidad de que todos los eventos enunciados más arriba sean ciertos es de ½ × ½ × 1/100 = 1/400. Se trata de una probabilidad bastante pequeña de que este esqueleto sea el de Ricardo III; los investigadores que llevaron a cabo originalmente este análisis asumieron una probabilidad previa «escéptica» de 1/40, así que nosotros estamos siendo considerablemente más escépticos.[153]

Pero cuando los arqueólogos examinaron el esqueleto con más detalle encontraron una notable serie de evidencia forense favorable al hallazgo, que incluía la datación por radiocarbono de los huesos (había un 95 % de probabilidades de que su origen estuviese entre 1456 y 1530), el hecho de que era un varón de alrededor de treinta años, el que el esqueleto presentase escoliosis (curvatura de la espina dorsal), y evidencia de que el cuerpo había sido mutilado después de la muerte. El análisis genético de los descendientes conocidos de familiares cercanos de Ricardo (él no tuvo hijos) reveló ADN mitocondrial compartido (por parte de su madre). El cromosoma masculino Y no apoyaba la existencia de relación, pero esto podía ser fácilmente explicado por rupturas en la línea masculina debido a paternidad errónea.

El valor de cada evidencia puede resumirse usando su razón de verosimilitud, que en esta situación se define como:

$$\text{razón de verosimilitud} = \frac{\text{probabilidad de la evidencia, si el esqueleto ES Ricardo III}}{\text{probabilidad de la evidencia, si el esqueleto NO ES Ricardo III}}$$

La tabla 11.1 muestra las razones de verosimilitud individuales para cada evidencia, y revela que ninguna es muy convincente individualmente, aunque los investigadores eran cautelosos y deliberadamente pecaron del lado de establecer unas razones de verosimilitud bajas, que no favorecían que el esqueleto fuese Ricardo III. Pero si asumimos que son descubrimientos forenses independientes, entonces estamos legitimados para multiplicar las razones de verosimilitud para obtener una evaluación global de la fuerza de la evidencia combinada, lo que resulta en un valor «extremadamente fuerte» de 6,7 millones. Los términos verbales empleados en la tabla están tomados de la escala mostrada en la tabla 11.2, que ha sido recomendada para su uso en los tribunales.[154]

Evidencia	Razón de verosimilitud (estimación conservadora)	Equivalente verbal
Datación por radiocarbono 1456-1530	1,8	Apoyo débil
Edad y sexo del esqueleto	5,3	Apoyo débil
Escoliosis	212	Apoyo moderadamente fuerte
Herida *post mortem*	42	Apoyo moderado
Coincidencia ADN mt	478	Apoyo moderadamente fuerte
Cromosoma Y no coincidente	0,16	Evidencia en contra débil
Evidencia combinada	6,7 millones	Apoyo más que extremadamente fuerte

Tabla 11.1. Razones de verosimilitud estimadas para evidencia encontrada en el esqueleto hallado en Leicester, comparando las hipótesis de que el esqueleto es, o no es, Ricardo III. La razón de verosimilitud combinada de 6,7 millones se obtiene al multiplicar juntas todas las razones de verosimilitud individuales (sin redondear).

Valor de la razón de verosimilitud	Equivalente verbal
1 – 10	Apoyo débil a la proposición
10 – 100	Apoyo moderado
100 – 1.000	Apoyo moderadamente fuerte
1.000 – 10.000	Apoyo fuerte
10.000 – 100.000	Muy fuerte
100.000 – 1.000.000	Extremadamente fuerte

Tabla 11.2. Interpretaciones verbales recomendadas de las razones de verosimilitud cuando se informa de descubrimientos forenses en los tribunales.

Entonces, ¿es convincente esta evidencia? Recordemos que hemos calculado una probabilidad inicial conservadora de 1 entre 400 de que este esqueleto sea Ricardo III, antes de tener en cuenta los detalles de los descubrimientos forenses. Esto corresponde a una razón de probabilidad inicial de alrededor de 1 entre 400: el teorema de Bayes nos dice que multipliquemos esto por la razón de verosimilitud para obtener una razón de verosimilitud final, que sería entonces 6,7 millones / 400 = 16.750. De manera que, incluso aunque seamos extremadamente cautelosos en estimar la razón de probabilidad inicial

y las razones de verosimilitud, podríamos decir que hay una razón de probabilidad de alrededor de 17.000 a 1 de que el esqueleto sea Ricardo III.

El propio análisis «escéptico» de los investigadores los llevó a una razón de probabilidad posterior de 167.000 a 1 o, lo que es lo mismo, una probabilidad de 0,999994 de que habían encontrado a Ricardo III. Esto fue considerado una evidencia suficiente para poder enterrar el esqueleto con todos los honores en la catedral de Leicester.

En casos legales, las razones de verosimilitud se vinculan normalmente a una evidencia de ADN que encuentra una coincidencia de algún grado entre el ADN del sospechoso y una muestra encontrada en la escena del crimen. Las dos hipótesis rivales son o que el sospechoso ha dejado ese rastro de ADN, o que sea de otra persona, de manera que podemos expresar la razón de verosimilitud de la manera siguiente:

$$\text{razón de verosimilitud} = \frac{\text{probabilidad de una coincidencia de ADN asumiendo que el rastro lo dejó el sospechoso}}{\text{probabilidad de una coincidencia de ADN, asumiendo que otro dejó el rastro}}$$

Normalmente se asume que el numerador en esta razón es 1, mientras que el denominador es la probabilidad de que una persona aleatoriamente escogida de entre la población hubiese proporcionado la coincidencia: esto se conoce como probabilidad de coincidencia aleatoria. Las razones de verosimilitud típicas para la evidencia de ADN pueden situarse en millones o miles de millones, aunque los valores exactos podrían ser discutidos, como cuando hay complicaciones debido a que las trazas contienen una mezcla de ADN proveniente de múltiples personas.

Los tribunales británicos permiten el uso de razones de verosimilitud individuales, pero no pueden multiplicarse, como en el caso de Ricardo III, dado que se supone que el proceso de combinar evidencias separadas es tarea del jurado.[155] Aparentemente, el sistema jurídico no está aún listo para adoptar la lógica científica.

¿Haría trampas en el póker el arzobispo de Canterbury?

Es un hecho poco conocido que el renombrado economista John Maynard Keynes estudió probabilidad, y desarrolló un experimento mental para ilustrar la importancia de tener en cuenta las razones de probabilidad iniciales cuando se evalúan las implicaciones de la evidencia. Nos pedía que imaginásemos que estábamos jugando al póker con el arzobispo de Canterbury, y que en la primera ronda, nos gana con una escalera real. ¿Deberíamos sospechar que está haciendo trampas?

La razón de verosimilitud para este evento es:

$$\text{razón de verosimilitud} = \frac{\text{probabilidad de escalera real, asumiendo que el arzobispo hace trampas}}{\text{probabilidad de escalera real, asumiendo que el arzobispo tiene suerte}}$$

Podríamos asumir que el numerador es 1, mientras que el denominador podría ser 1/72.000, lo que nos daría una razón de verosimilitud de 72.000; utilizando los estándares de la tabla 11.2, esto correspondería a la evidencia «muy sólida» de que el arzobispo está haciendo trampas. Aun así, ¿deberíamos concluir que realmente está haciendo trampas? El teorema de Bayes nos dice que nuestra razón de probabilidad final debe basarse en el producto de esta razón de verosimilitud por la razón de probabilidad inicial. Parece razonable asumir que, al menos, antes de que hayamos empezado a jugar, nuestra razón de probabilidad acerca de que el arzobispo esté haciendo trampas sería baja, quizá de 1 a 1.000.000, dado que se supone que es un clérigo respetable. De manera que el producto de la razón de verosimilitud y la razón previa de probabilidad sería de alrededor de 72.000/1.000.000, lo que nos da una razón de probabilidad de alrededor de 7/100, lo que, a su vez, corresponde a una probabilidad de 7/107, o un 7 % de que sea un tramposo. De manera que, en principio, le damos el beneficio de la duda, aunque probablemente no seríamos tan generosos con alguien que acabásemos de conocer en un pub. Y quizá deberíamos vigilar estrechamente al arzobispo.

Inferencia estadística bayesiana

El teorema de Bayes, aunque no esté permitido en los tribunales del

Reino Unido, es una forma científicamente correcta de cambiar nuestra opinión sobre la base de evidencia nueva. Las frecuencias esperadas hacen que el análisis bayesiano sea razonablemente claro para situaciones simples que implican solo dos hipótesis, digamos, por ejemplo, sobre si alguien tiene o no una enfermedad, o ha cometido o no un delito. No obstante, las cosas se complican cuando queremos aplicar las mismas ideas a inferencias sobre cantidades desconocidas que podrían adoptar todo un rango de valores, como por ejemplo los parámetros en los modelos estadísticos.

El objetivo del artículo original de 1763 del reverendo Thomas Bayes era responder a una pregunta muy básica de ese tipo: dado que algo ha pasado o no ha pasado en un número conocido de ocasiones similares, ¿qué probabilidad deberíamos asignar a que ocurra la próxima vez?[156] Por ejemplo, si lanzamos una chincheta 20 veces y sale con la punta hacia arriba 15 veces y hacia abajo 5, ¿cuál es la probabilidad de que salga con la punta hacia arriba la vez siguiente? A estas alturas, podría usted pensar que la respuesta es obvia: 15/20 = 75 %. Pero esta podría no ser la respuesta del reverendo: podría decir más bien 16/22 = 73 %. ¿Cómo habría llegado a esta conclusión?

Bayes usó como metáfora una mesa de billar[157] que no podemos ver. Supongamos que se lanza a la mesa una bola blanca de manera aleatoria, se marca su posición en la mesa con una línea y luego se saca la bola blanca. A continuación se arrojan aleatoriamente a la mesa un número de bolas rojas, y solamente se le informa de cuántas han caído a la izquierda y cuántas a la derecha de la línea. ¿Dónde piensa usted que estaría la línea, y cuál sería la probabilidad de que la siguiente bola roja cayese a la izquierda de la misma?

Digamos, por ejemplo, que se arrojan cinco bolas rojas, y se nos dice que dos han caído a la izquierda y tres a la derecha de la línea dejada por la bola blanca, como se muestra en la figura 11.4(a). Bayes mostró que nuestras creencias sobre la posición de la línea deberían ser descritas por la distribución de probabilidad mostrada en la figura 11.4(b) —la matemática es bastante compleja, se puede ver en la nota a pie de página—.[158] La posición de la línea discontinua, que indica dónde ha caído la bola blanca, se estima que es 3/7 en toda la mesa, que es la media (esperanza) de esta distribución.

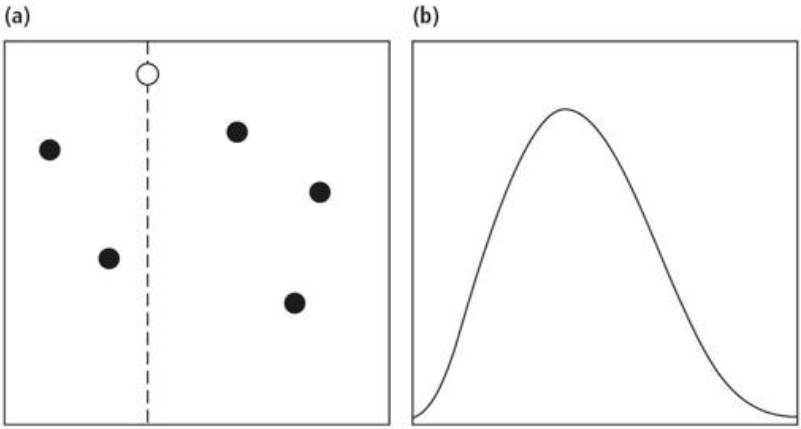


Figura 11.4. Mesa de «billar» de Bayes. (a) Se lanza una bola blanca a la mesa y la línea discontinua indica la posición en la que ha caído. Se lanzan cinco bolas rojas a la mesa y caen como se muestra. (b) Un observador no puede ver la mesa, y solo se le dice que dos bolas rojas han caído a la izquierda, y tres a la derecha de la línea discontinua. La curva representa la distribución de probabilidad para el observador acerca de dónde ha caído la bola blanca, superpuesta sobre la mesa de billar. La media de la curva es 3/7, que es también la actual probabilidad del observador sobre que la siguiente bola roja caiga a la izquierda de la línea.

Este valor de 3/7 podría parecer raro, dado que el estimador intuitivo sería más bien 2/5 —la proporción de bolas rojas que han caído a la izquierda de la línea—. Pero Bayes mostró que en estas circunstancias deberíamos estimar la posición como:

caído a la izquierda de la línea—. Pero Bayes mostró que en estas circunstancias deberíamos estimar la posición como:

$$\frac{\text{Número de bolas rojas que han caído a la izquierda} + 1}{\text{Número total de bolas rojas} + 2}$$

Esto significa, por ejemplo, que antes de que se lance ninguna bola roja, podemos estimar que la posición sería $(0 + 1)/(0 + 2) = 1/2$, mientras que el enfoque intuitivo sugeriría más bien que no podríamos dar ninguna respuesta, dado que aún no hay datos. Lo que está haciendo Bayes es hacer uso de la información acerca de cómo se ha decidido inicialmente la posición de la línea, dado que sabemos que es escogida aleatoriamente arrojando la bola blanca. Esta información

inicial tiene el mismo papel que la prevalencia usada en las pruebas de cáncer de mama o de dopaje —se conoce como información previa e influye en nuestras conclusiones finales—. De hecho, dado que la fórmula de Bayes añade un 1 al número de bolas rojas a la izquierda de la línea y un 2 al número total de bolas rojas, podríamos pensar que es equivalente a haber arrojado dos bolas rojas «imaginarias», y que una hubiese acabado a cada lado de la línea discontinua.

Nótese que si ninguna de las cinco bolas hubiese caído a la izquierda de la línea discontinua, no habríamos estimado su posición como 0/5, sino más bien como 1/7, que parece mucho más razonable. El estimador de Bayes nunca puede ser 0 o 1, y siempre está más cerca de ½ que la proporción simple: esto se conoce como contracción, en el sentido de que los estimadores siempre se comprimen o encogen hacia el centro de la distribución inicial, en este caso ½.

El análisis bayesiano utiliza la información sobre cómo se decidió la posición de la línea discontinua para establecer una distribución previa para esta posición, combinada con evidencia de los datos conocida como verosimilitud, para darnos una conclusión final denominada distribución posterior, que expresa todo lo que creemos actualmente sobre la cantidad desconocida. Así, por ejemplo, el software informático puede calcular que un intervalo que va de 0,12 a 0,78 contiene el 95 % de la probabilidad de la figura 11.4(b), de manera que podemos afirmar con una seguridad del 95 % que la línea que marca la bola blanca está entre estos límites. Este intervalo se estrechará cada vez más a medida que se lanzan más y más bolas rojas a la mesa y se anuncian sus posiciones con relación a la línea, hasta que finalmente convergeremos a la respuesta correcta.

La principal controversia acerca del análisis bayesiano es la fuente de la distribución previa. En la mesa de billar de Bayes, la bola blanca se arroja de manera aleatoria sobre la mesa, por lo que todo el mundo estaría de acuerdo en que la distribución previa está uniformemente dispersa a lo largo de toda la línea entre 0 y 1. Cuando no contamos con este tipo de conocimiento físico, se pueden hacer distintas sugerencias para obtener distribuciones previas, incluyendo el uso de juicios subjetivos, el aprendizaje a partir de datos históricos y la especificación de distribuciones previas objetivas que intenten que los datos hablen por sí mismos sin introducir juicios subjetivos.

Quizá la idea más importante que extraer de todo esto es que no hay una distribución previa «verdadera», y que cualquier análisis debería incluir un estudio de sensibilidad para un número de opciones alternativas, que comprendiesen toda una serie de opiniones posibles

diferentes.

¿Cómo podemos hacer un mejor análisis de las encuestas preelectorales?

Hemos visto cómo el análisis bayesiano proporciona un mecanismo formal que usa el conocimiento contextual para hacer inferencias más realistas sobre el problema particular al que nos enfrentamos. Estas ideas pueden (literalmente) llevarse a otro nivel mediante modelos jerárquicos o multinivel que analizan simultáneamente varias cantidades individuales: el poder de esos modelos se puede ver en su éxito cuando se aplican a las encuestas preelectorales.

Sabemos que las encuestas deberían basarse idealmente en muestras grandes, aleatorias y representativas, pero este tipo de encuestas son cada vez más caras, y, en cualquier caso, la gente está cada vez menos dispuesta a responder a encuestas. Por ello, las empresas demoscópicas cada vez se basan más en paneles online, que sabemos que no son realmente representativos, pero sobre los que se emplean sofisticados modelos estadísticos para intentar determinar cuáles hubiesen sido las respuestas si las empresas hubieran podido emplear una muestra aleatoria bien hecha. Esto nos recuerda aquel dicho de que hay que tener cuidado si intentas hacer un bolso de seda a partir de una oreja de cerdo.

Las cosas empeoran aún más en el caso de las encuestas preelectorales, dado que las actitudes podrían no ser uniformes a lo largo de todo el país, y, por ello, afirmaciones acerca de todo el país podrían derivarse de agregar resultados de estados o electorados muy distintos. Idealmente, necesitaríamos sacar conclusiones a nivel local, pero los participantes en el panel online estarán distribuidos de manera no aleatoria a lo largo de estas áreas locales, lo que quiere decir que hay datos muy limitados sobre los que basar los análisis locales.

La respuesta bayesiana a este problema se conoce como regresión multinivel y post-estratificación (RMP). La idea básica es dividir a todos los posibles votantes en pequeñas «celdas», cada una de las cuales comprende un grupo de personas muy homogéneo —que viven en la misma área, con la misma edad, género, comportamiento electoral pasado, y otras características medibles—. Podemos usar datos demográficos para estimar el número de personas en cada celda, cada una de las cuales se asume que tiene la misma probabilidad de votar por un determinado partido. El problema es establecer cuál es esa probabilidad, cuando nuestros datos no aleatorios podrían significar que solo tenemos unas pocas personas, o quizás ninguna, en

una determinada celda.

El primer paso es construir un modelo de regresión para la probabilidad de votar de una forma particular dadas las características de la celda, de manera que nuestro problema se reduce a estimar los coeficientes de la ecuación de regresión. Pero todavía hay demasiados coeficientes para realizar una estimación fiable con los métodos estándar, y aquí es donde entran en juego las ideas bayesianas. Se asume que los coeficientes que corresponden a distintas áreas son similares, un cierto punto intermedio entre asumir que son exactamente los mismos y asumir que están completamente no relacionados.

Matemáticamente, se puede demostrar que esta asunción es equivalente a asumir que estas cantidades desconocidas han sido extraídas de la misma distribución previa, y esto nos permite acercar muchos estimadores individuales bastante imprecisos, lo que da lugar a conclusiones más finas, más fiables, que no estén tan influidas por unas pocas observaciones atípicas. Una vez realizados estos estimadores más sólidos del comportamiento electoral dentro de cada una de las miles de celdas, los resultados pueden ser combinados para generar una predicción sobre cómo votará todo el país.

En las elecciones presidenciales de Estados Unidos de 2016, las encuestas basadas en regresión multinivel y post-estratificación acertaron en 50 de los 51 estados y el distrito de Columbia, y solo se equivocaron en Míchigan, sobre la base de entrevistas a 9.485 votantes en las semanas que precedieron a las elecciones. Unas predicciones igualmente buenas se realizaron para las elecciones de 2017 en el Reino Unido, donde la empresa demoscópica YouGov entrevistó a 50.000 personas a la semana sin preocuparse de que fuesen una muestra representativa. Después aplicaron la RMP para predecir un parlamento sin mayoría absoluta, y donde los conservadores obtendrían un 42 % de los votos, que es exactamente lo que pasó. Las encuestas que usaban métodos más tradicionales fallaron estrepitosamente.[159]

Así que ¿podemos realmente hacernos el bolso de seda a partir de una oreja no aleatoria de cerdo? La RMP no es una panacea: si un gran número de entrevistados ofrecen respuestas sistemáticamente engañosas y, por tanto, no representan a su «celda», entonces no hay análisis estadístico, por sofisticado que sea, que pueda contrarrestar ese sesgo. Pero parece ser beneficioso usar modelos bayesianos, y veremos más adelante si esto ha tenido un éxito espectacular en las encuestas a pie de urna llevadas a cabo el día de las elecciones.

El «pulido» bayesiano puede conferir precisión cuando se tienen pocos datos. Es una técnica que se usa cada vez más para modelizar, por ejemplo, cómo se expanden las enfermedades en el espacio y en el tiempo. El aprendizaje bayesiano ahora se ve como un proceso fundamental para determinar cómo el individuo se hace consciente del entorno. Partimos de unas expectativas previas sobre lo que nos vamos a encontrar en cualquier contexto, y después solo tenemos que tomar nota de características inesperadas, que se usan después para actualizar nuestra percepción actual. Esta es la idea detrás de la llamada «mente bayesiana».[160]

Los mismos procedimientos de aprendizaje han sido implementados en los coches automáticos, que tienen un «mapa mental» probabilístico del entorno que los rodea que se actualiza constantemente para que reconozcan las luces de tráfico, la gente, otros coches y demás: «Esencialmente, un coche robot "piensa" en sí mismo como una masa de probabilidades, que viaja por una carretera bayesiana».[161]

Estos problemas se refieren a la estimación de cantidades que describen el mundo, pero el uso de métodos bayesianos para evaluar hipótesis científicas sigue siendo más controvertido. Al igual que en las pruebas de Neyman-Pearson, necesitamos primero establecer dos hipótesis rivales. Una hipótesis nula H0, que es normalmente la ausencia de algo, como que no haya bosón de Higgs, o un tratamiento médico que no tenga efecto. La hipótesis alternativa, H1, dice que existe algo importante.

Las ideas que hay detrás de la prueba de hipótesis bayesiana son, por tanto, esencialmente las mismas que en los casos legales, en los cuales la hipótesis nula es generalmente la inocencia, la alternativa es la culpabilidad, y expresamos el apoyo relativo que una determinada evidencia proporciona a estas dos hipótesis por medio de la razón de verosimilitud. Para la prueba científica de hipótesis, el equivalente preciso de la razón de verosimilitud es el factor de Bayes, con la diferencia de que las hipótesis científicas generalmente contienen parámetros desconocidos, como por ejemplo el efecto real bajo la hipótesis alternativa. El factor de Bayes solo puede obtenerse haciendo una media con respecto a la distribución previa de los parámetros desconocidos, lo que hace que la distribución previa —la parte más controvertida de un análisis bayesiano— sea crucialmente importante. Por ello, los intentos de reemplazar las pruebas de significatividad por factores de Bayes, particularmente en psicología, son una fuente considerable de conflicto. Los críticos argumentan que detrás de cualquier factor de Bayes asoma el supuesto de una distribución previa para cualquier parámetro desconocido tanto en la hipótesis

nula como en la alternativa.

Robert Kass y Adrian Raftery son dos renombrados estadísticos bayesianos que propusieron una escala ampliamente utilizada de factores de Bayes que se muestra en la tabla 11.3. Nótese el contraste con la tabla 11.2 con respecto a la interpretación verbal de razones de probabilidad para los casos legales, en los que se exigía una razón de probabilidad de 10.000 para declarar que la evidencia era «muy robusta», frente a hipótesis científicas que solo necesitan un factor de Bayes mayor de 150. Esto refleja quizás la necesidad de establecer la culpabilidad en casos penales «más allá de toda duda razonable», mientras que las afirmaciones científicas se basan en una evidencia más débil, y muchas son rechazadas por investigaciones posteriores.

Factor de Bayes	Solidez de la evidencia
1 a 3	Apenas vale la pena mencionarla
3 a 20	Positiva
20 a 150	Fuerte
>150	Muy fuerte

Tabla 11.3. Escala de Kass y Raftery para interpretar los factores de Bayes en favor de una hipótesis. (R. E. Kass y A. E. Raftery, «Bayes Factors», Journal of the American Statistical Association 90 (1995), pp. 773-795).

En nuestro capítulo sobre pruebas de hipótesis se afirmaba que un valor P de 0,05 suponía una «evidencia débil». El razonamiento está en parte basado en factores de Bayes: P = 0,05 correspondería, bajo algunas hipótesis alternativas, a factores de Bayes entre 2,4 y 3,4, lo que se sugiere en la tabla 11.3 que es evidencia más bien débil. Como vimos en el capítulo 10, esto llevó a proponer una reducción del valor P necesario para afirmar que se había realizado un «descubrimiento» a 0,005.

A diferencia de los test de significatividad de la hipótesis nula, los factores de Bayes tratan a las dos hipótesis simétricamente, y, por tanto, pueden sustentar una hipótesis nula. Y si queremos asignar probabilidades previas a hipótesis, podríamos incluso calcular las probabilidades posteriores de teorías alternativas sobre cómo funciona el mundo. Supongamos, basándonos exclusivamente en la teoría, que pensamos que hay una probabilidad de 50:50 de que exista el bosón de Higgs, lo que correspondería a una razón de probabilidad de 1. Los datos que vimos en el anterior capítulo nos dan un valor P de alrededor de 1/3.500.000, lo que puede convertirse en un factor de

Bayes máximo de alrededor de 80.000 a favor del bosón de Higgs, lo que sin duda es una evidencia muy contundente, incluso de acuerdo con los usos jurídicos.

Cuando se combina con unas razones previas de probabilidad de 1, el resultado son unas razones de probabilidad posteriores de 0,99999. Pero ni la comunidad jurídica ni la científica aprueban normalmente este tipo de análisis, aunque pueda usarse para el caso de Ricardo III.

Una batalla ideológica

En este libro hemos ido del examen informal de los datos, pasando por la comunicación de estos mediante estadísticos de resumen, al uso de los modelos de probabilidad para llegar a intervalos de confianza, valores P, *etc.* Estas herramientas estándar para hacer inferencias, con las que han batallado generaciones de estudiantes, se conocen como métodos «clásicos» o «de frecuencias», ya que se basan en las propiedades muestrales de la estadística.

El enfoque bayesiano alternativo se basa en principios fundamentalmente diferentes. Como hemos visto, la evidencia externa sobre magnitudes desconocidas, expresada como una distribución previa, se combina con la evidencia de un modelo de probabilidad subyacente de los datos, conocido como verosimilitud, para llegar a una distribución posterior final que forma la base de todas las conclusiones.

Si adoptamos seriamente esta filosofía estadística, las propiedades muestrales de la estadística se vuelven irrelevantes. De manera que después de pasar años aprendiendo que un intervalo de confianza del 95 % no significa que haya una probabilidad del 95 % de que el valor real esté en el intervalo,[162] el pobre estudiante tiene que olvidarse de todo eso: un intervalo bayesiano de incertidumbre del 95 % tiene precisamente esa interpretación.

Pero el argumento sobre la forma «correcta» de llevar a cabo la inferencia estadística es aún más complejo que una simple disputa entre frecuentistas y bayesianos. Al igual que en el caso de los movimientos políticos, cada escuela se divide en múltiples facciones que a menudo están en conflicto unas con otras.

En la década de 1930, estalló un conflicto a tres bandas en el debate público. Tuvo lugar en la Real Sociedad Estadística, que entonces como ahora recogía y publicaba meticulosamente las discusiones en torno a trabajos presentados en sus reuniones. En 1934 Jerzy Neyman

propuso su teoría de los intervalos de confianza; Arthur Bowley, un firme defensor del método del enfoque bayesiano, entonces conocido como probabilidad inversa, dijo: «No estoy nada seguro de que la "confianza" no sea un "abuso de confianza"», y añadió que sería necesario adoptar un enfoque bayesiano: «¿Realmente esto nos hace avanzar? […] ¿Nos lleva realmente hacia lo que necesitamos: la probabilidad de que en el universo del que estamos sacando una muestra la proporción esté dentro […] de ciertos límites? Pienso que no». La asociación socarrona de los intervalos de confianza con los abusos de confianza continuaría en las décadas subsiguientes.

Al año siguiente, en 1935, se desató la guerra abierta entre los dos campos no bayesianos, con Ronald Fisher en un lado, y Jerzy Neyman y Egon Pearson en el otro. El enfoque de Fisher se basaba en la estimación usando la función de «verosimilitud», que expresaba el apoyo relativo que otorgaban los datos a los diferentes parámetros, y las pruebas de hipótesis basadas en los valores P. Por el contrario, el enfoque Neyman-Pearson, que, como hemos visto, se conocía como comportamiento inductivo, se centraba en gran medida en la toma de decisiones: si se decide que la respuesta correcta está en un intervalo de confianza del 95 %, entonces se tendrá razón el 95 % del tiempo, y se deberían controlar los errores de tipo I y de tipo II cuando se comprueban hipótesis. Sugirieron incluso que se debería «aceptar» la hipótesis nula cuando estuviese incluida en el intervalo de confianza del 95 %, un concepto que era anatema para Fisher (y que posteriormente fue rechazado por la comunidad estadística).

Fisher primero acusó a Neyman de «caer en la serie de malentendidos que revela su artículo». Entonces, Pearson salió en defensa de Neyman, diciendo que «aunque sabíamos que había una creencia muy extendida en la infalibilidad del profesor Fisher, debería, en primer lugar, pensar si es muy sensato acusar a un compañero de incompetencia sin, al mismo tiempo, demostrar que ha resuelto con éxito la cuestión objeto de disputa». La agria discusión entre Fisher y Neyman duraría décadas.

La lucha por la supremacía ideológica en el mundo de la estadística continuaría tras la Segunda Guerra Mundial, pero a lo largo del tiempo las escuelas más estándar, no bayesianas, han llegado a un compromiso pragmático, con experimentos que normalmente se diseñan usando un enfoque Neyman-Pearson de errores de tipo I y de tipo II, pero después analizados desde una perspectiva fisheriana, usando valores P como medida de evidencia. Como hemos visto en el contexto de los ensayos clínicos, esta extraña amalgama parece funcionar bastante bien, llevando al prominente estadístico

(bayesiano) Jerome Cornfield a afirmar que «la paradoja es que a pesar de todo ha emergido una estructura sólida de valores permanentes, a la que le falta solamente el firme fundamento lógico sobre el que originalmente se pensaba que tenía que haber sido construida».[163]

Las supuestas ventajas de los métodos estadísticos convencionales sobre el bayesianismo son la aparente separación entre la evidencia de los datos y cualquier factor subjetivo, la facilidad de la programación, la amplia aceptación de los criterios de «significatividad», la disponibilidad de software y la existencia de métodos sólidos que no tienen que hacer asunciones fuertes sobre la forma de la distribución. Por otro lado, los entusiastas del enfoque bayesiano afirmarían que la capacidad de usar elementos externos, e incluso de elementos explícitamente subjetivos, es lo que permite realizar inferencias y predicciones más potentes.

La comunidad estadística entablaba a menudo largas y agrias polémicas sobre los fundamentos de la materia, pero ahora se ha establecido una tregua y la norma es adoptar un enfoque más global, según el cual los métodos se escogen de acuerdo con el contexto práctico, y no tanto según las credenciales ideológicas derivadas de Fisher, Neyman-Pearson o Bayes. Este parece una postura sensata y pragmática para una discusión que puede parecer un tanto oscura para los no estadísticos. Mi punto de vista personal es que, aunque podrían estar en desacuerdo sobre los elementos fundamentales de su materia, unos estadísticos razonables normalmente llegarán a conclusiones similares. Los problemas que surgen en la ciencia estadística no provienen generalmente de la filosofía subyacente a los métodos precisos que usa. En lugar de ello, es más probable que se deban a un diseño inadecuado, datos sesgados, asunciones inapropiadas y, lo que quizá es más importante, una mala práctica científica. En el capítulo siguiente echaremos un vistazo a este lado oscuro de la estadística.[164]

Resumen

→ Los métodos bayesianos combinan la evidencia de los datos (resumida por la verosimilitud) con creencias iniciales (conocidas como distribución previa) para generar una probabilidad posterior para la magnitud desconocida.

→ El teorema de Bayes para dos hipótesis rivales puede ser expresado como razón de probabilidad posterior = razón de verosimilitud × razón previa de probabilidad.

→ La razón de verosimilitud expresa el apoyo relativo a dos hipótesis que proporciona una determinada evidencia, y se usa normalmente para resumir la evidencia forense en los casos penales.

→ Cuando la distribución previa proviene de algún proceso físico de muestreo, los métodos bayesianos no generan controversia. No obstante, normalmente se necesita ser juicioso.

→ Los modelos jerárquicos permiten combinar la evidencia a lo largo de múltiples análisis menores que se asume que tienen parámetros en común.

→ Los factores de Bayes son el equivalente de las razones de verosimilitud para las hipótesis científicas, y son un sustituto controvertido para los test de significatividad de la hipótesis nula.

→ La teoría de la inferencia estadística tiene una larga historia de controversias, pero son más importantes las cuestiones relacionadas con la calidad de los datos y la seriedad científica.

[150] Murió sin ser consciente de la longevidad de su legado. No solo su artículo seminal fue publicado póstumamente, en 1763, sino que su nombre no se asoció con este enfoque hasta el siglo XX.

[151] Algunos dirían incluso que fui adoctrinado.

[152] Una razón de probabilidad de 1 se conoce normalmente como «igualada», dado que los eventos son igualmente probables, o igualmente equilibrados.

[153] T. E. King et al., «Identification of the Remains of King Richard III», Nature Communications 5 (2014), p. 5631.

[154] Se pueden encontrar consejos sobre cómo comunicar las razones de probabilidad en: http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

[155] Para un artículo sobre el uso de Bayes en los tribunales, véase «A Formula for Justice», The Guardian, 2 de octubre de 2011.

[156] Sus palabras exactas fueron: «Dado el número de veces en las cuales un evento desconocido ha ocurrido o no: es obligatorio que el azar de la probabilidad de que ocurra en una única vez esté entre cualesquiera dos grados de probabilidad que puedan ser señalados». Son razonablemente claras, excepto que en la terminología moderna

probablemente invertiríamos su uso de «azar» y «probabilidad».

[157] Siendo como era un ministro presbiteriano, la llamó simplemente «mesa».

[158] La fórmula para esta distribución es 60p2 (1 − p)3, lo que se conoce técnicamente como distribución beta (3, 4). Con una distribución previa uniforme, la distribución posterior para la posición de la bola blanca, habiendo lanzado n bolas rojas y habiendo caído r a la izquierda de la blanca, es ((n + 1)! / r!(n − r)!) pr (1 − p)n−r, que es una distribución beta (r + 1, n − r +1).

[159] D. K. Park, A. Gelman y J. Bafumi, «Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls», Political Analysis 12 (2004), pp. 375-385; los resultados de YouGov aparecen en: https://yougov.co.uk/news/2017/06/14/how-we-correctly-called-hung-parliament/.

[160] K. Friston, «The History of the Future of the Bayesian Brain», Neuroimage 62:2 (2012), pp. 1230-1233.

[161] N. Polson y J. Scott, AIQ: How Artificial Intelligence Works and How We Can Harness Its Power for a Better World, Penguin, 2018, p. 92.

[162] Recuérdese que esto significa que, a largo plazo, el 95 % de esos intervalos contendrán el valor real, pero no podemos decir nada sobre un intervalo en particular.

[163] J. Cornfield, «Sequential Trials, Sequential Analysis and the Likelihood Principle», American Statistician 20 (1966), pp. 18-23.

[164] Pero yo sigo prefiriendo el enfoque bayesiano.

12

Cuando las cosas se hacen mal

¿Existe la percepción extrasensorial (PES)?

En 2011, el eminente psicólogo social estadounidense Daryl Bem publicó un artículo fundamental en una prominente revista de psicología que incluía el siguiente experimento. Un centenar de estudiantes se sienta enfrente de una pantalla de ordenador que muestra dos cortinas, y escogen cuál de ellas, la de la izquierda o la de la derecha, oculta una imagen. Después se «abren» las cortinas para revelar si han acertado o no, y esto se repite para una serie de 36 imágenes. El truco es que, sin que lo supieran los participantes, la posición de la imagen era determinada de manera aleatoria después de que el sujeto hubiese hecho su elección, de manera que cualquier exceso de decisiones correctas por encima del puro azar sería adscrito a una precognición acerca de dónde aparecería la imagen.

Bem señaló que, en lugar de la tasa de éxito esperada de un 50 % bajo la hipótesis nula de que no hay precognición, los sujetos escogieron correctamente un 53 % de las veces cuando se mostraba una imagen erótica ($P = 0{,}01$). El artículo contenía los resultados de ocho experimentos adicionales en precognición, con alrededor de mil participantes a lo largo de diez años, y se observaban resultados estadísticamente significativos en favor de la precognición en ocho de los nueve estudios. ¿Es esta una prueba convincente de que existe la percepción extrasensorial (PES)?

Este libro, espero, ha ilustrado algunas aplicaciones muy importantes de la ciencia estadística para la resolución de problemas del mundo real, llevadas a cabo con habilidad y cautela por parte de analistas que son conscientes de sus limitaciones y dificultades. Pero el mundo real no siempre es tan digno de admiración. Es hora de ver lo que ocurre cuando la ciencia y el arte de la estadística no se aplican con tanta escrupulosidad. Después veremos cómo fue recibido y criticado el artículo de Bem.

Hay una razón por la cual se está dedicando tanta atención a las malas prácticas estadísticas: se las considera culpables de lo que se conoce como crisis de reproductibilidad en la ciencia.

La «crisis de reproductibilidad»

El capítulo 10 exploró la notable afirmación de John Ioannidis en 2005 de que la mayor parte de los resultados científicos publicados eran falsos. Desde entonces, muchos otros investigadores han argumentado que hay una falta fundamental de fiabilidad en la literatura científica publicada. Los científicos no han podido replicar estudios realizados por sus pares, lo que sugiere que los estudios originales no son tan dignos de confianza como se pensaba previamente. Aunque inicialmente se centraron en la medicina y la biología, estas acusaciones se han extendido desde entonces a la psicología y a otras ciencias sociales, aunque el porcentaje real de afirmaciones que son exageradas o falsas es objeto de discusión.

La alegación original de Ioannidis se basaba en un modelo teórico, pero un enfoque alternativo consistiría en coger estudios pasados e intentar replicarlos, en el sentido de llevar a cabo experimentos similares y ver si se observan resultados similares. El Proyecto de Reproductibilidad fue una gran empresa colaborativa en la cual cien estudios de psicología fueron repetidos usando muestras mayores, y, por tanto, una potencia mayor para poder detectar un efecto real, si es que existía. El proyecto reveló que aunque un 97 % de los estudios originales tuvieron resultados estadísticamente significativos, solo los mantuvieron el 36 % de las réplicas.[165]

Desgraciadamente, esto fue ampliamente presentado como que el otro 63 % de estudios «significativos» eran afirmaciones falsas, pero esto cae en la falacia de la división estricta entre estudios que son o significativos o no significativos. El distinguido estadístico y bloguero estadounidense Andrew Gelman señaló que la «diferencia entre "significativo" y "no significativo" no es estadísticamente significativa».[166] De hecho, solo el 23 % de los estudios originales y los replicados tuvieron resultados diferentes unos de otros, lo que es quizá una estimación más apropiada del porcentaje de estudios originales con afirmaciones exageradas o falsas.

Más que pensar en términos de significatividad o no significatividad como determinantes de un «descubrimiento», sería mejor centrarse en el tamaño de los efectos estimados. El Proyecto de Reproductibilidad descubrió que los efectos de la replicación iban, de media, en la misma dirección que los estudios originales, pero su magnitud era aproximadamente la mitad. Esto apunta a un importante sesgo en la literatura científica; un estudio que ha descubierto algo «grande», de lo que al menos parte puede considerarse pura suerte, es probable que lleve a una prominente publicación. En una analogía con la regresión a la media, esto podría denominarse «regresión a la hipótesis nula», porque unas estimaciones exageradas de los efectos disminuyen más

adelante hasta acercarse a la hipótesis nula.

La supuesta crisis de reproductibilidad es una cuestión compleja, basada en la excesiva presión a la que están sometidos los investigadores para hacer «descubrimientos» y publicar sus resultados en prestigiosas revistas científicas, todo lo cual depende crucialmente de encontrar resultados estadísticamente significativos. Ninguna institución o profesión concreta es la culpable. También hemos mostrado cuando discutíamos las pruebas de hipótesis que, incluso si la práctica estadística fuese perfecta, unos efectos verdaderos y sustanciales son tan raros que una importante proporción de resultados que se consideran «significativos» son, inevitablemente, falsos positivos (figura 10.5). Pero, como veremos a continuación, la práctica estadística en muchos casos está lejos de ser perfecta.

La estadística se puede hacer mal en cada estadio del ciclo PPDAC. Desde el comienzo, podemos intentar resolver un Problema que simplemente no puede responderse con la información disponible. Por ejemplo, si nos proponemos explicar por qué las tasas de embarazos adolescentes han caído tan acusadamente en el Reino Unido en la última década, no encontraremos nada en los datos observados que pueda ofrecer una explicación.[167]

El Plan también puede ir mal encaminado, por ejemplo por las siguientes razones:

→ Escoger una muestra que sea conveniente y barata pero no representativa, por ejemplo las encuestas preelectorales telefónicas.

→ Hacer preguntas capciosas en las encuestas, o con una redacción que lleve a confusión, como, por ejemplo: «¿Cuánto piensa usted que puede ahorrar comprando en Internet?».

→ No hacer una comparación justa, como por ejemplo evaluar la homeopatía observando solamente a los voluntarios de la terapia.

→ Diseñar un estudio demasiado pequeño, y, por tanto, con poca potencia, lo que significa que se detectarán menos hipótesis alternativas correctas.

→ No recoger datos de variables de confusión potenciales, falta de enmascaramiento en los ensayos aleatorizados, *etc.*

Como dijo célebremente Ronald Fisher, «consultar al estadístico después de que haya terminado un experimento es muchas veces

pedirle que lleve a cabo una autopsia. Quizá pueda decirnos de qué murió el experimento».[168]

Respecto al paso de reunir Datos, algunos problemas comunes son, por ejemplo, demasiados valores perdidos, personas que abandonan el estudio, un reclutamiento mucho más lento de lo esperado o, simplemente, no codificar todo eficientemente. Todas estas cuestiones deberían haber sido previstas y evitadas por medio de un cuidadoso ensayo previo.

La forma más fácil de que el Análisis vaya mal es simplemente cometer un error. Muchos de nosotros habremos cometido errores de codificación o en las hojas de cálculo, pero quizás no con las consecuencias de los siguientes ejemplos:

→ Los destacados economistas Carmen Reinhart y Kenneth Rogoff publicaron un artículo en 2010 que influyó fuertemente en las actitudes hacia la austeridad. Un estudiante de doctorado descubriría más adelante que cinco países habían quedado fuera del análisis principal sin querer debido a un sencillo error en la hoja de cálculo.[169]

→ Un programador de AXA Rosenberg, una empresa global de inversiones inmobiliarias, programó incorrectamente un modelo estadístico de tal manera que algunos de sus factores de riesgo eran demasiado pequeños por un factor de 10.000, lo que llevó a 217 millones de pérdidas para sus clientes. En 2011 la Comisión de Bolsa y Valores (SEC)[170] multó a AXA Rosenberg por esta cantidad más unos 25 millones de dólares adicionales en penalizaciones, lo que en total supuso una multa de 242 millones de dólares por no informar a los inversores de un error de riesgos en un modelo.[171]

En ocasiones los cálculos son computacionalmente correctos, pero emplean métodos estadísticos incorrectos. Algunos candidatos habituales al método más inapropiado son los siguientes:

→ Llevar a cabo un ensayo «aleatorizado por conglomerados» en el que grupos de personas, como por ejemplo todos los pacientes de medicina general, han sido asignados aleatoriamente de manera simultánea a una intervención particular, y aun así analizar los datos como si la gente hubiese sido aleatorizada individualmente.

→ Medir dos grupos antes y después de una intervención, y considerar que los grupos son diferentes si uno ha cambiado significativamente, y el cambio del otro grupo no es significativo. El

procedimiento correcto es llevar a cabo una prueba estadística formal para comprobar si los grupos son diferentes (lo que se conoce como una prueba de interacción).

→ Interpretar la «no significatividad» como sinónimo de «no efecto». Por ejemplo, en el estudio sobre alcohol y mortalidad mencionado en el capítulo 10, hombres de entre 50 y 64 años con un consumo de 15-20 unidades de alcohol a la semana tenían asociado un riesgo de mortalidad significativamente reducido, mientras que la reducción en hombres que beben un poco menos o un poco más no es significativamente distinta de cero. El artículo argumentaba que era una diferencia importante, pero los intervalos de confianza revelaban que había una diferencia insignificante entre estos grupos. Una vez más, la diferencia entre significativo y no significativo no es necesariamente significativa.

Cuando se trata de redactar Conclusiones, quizá la forma más flagrante de mala práctica es llevar a cabo muchas pruebas estadísticas y después informar únicamente de las más significativas e interpretarlas al pie de la letra. Hemos visto que esto incrementa enormemente la probabilidad de encontrar un valor P significativo, hasta el punto de revivir a un pez muerto. Es el equivalente de televisar únicamente los goles que marca un equipo, y no los goles que le meten: es imposible obtener una impresión certera con una forma de informar tan selectiva.

La información selectiva de los resultados es algo que empieza a cruzar la frontera entre la simple incompetencia y la falta de ética, y, de hecho, existe evidencia inquietante de que esto no es nada infrecuente. En Estados Unidos ha habido incluso una condena penal por informar selectivamente de resultados significativos en un análisis de subconjuntos. Scott Harkonen era el director general de InterMune, una empresa que llevó a cabo un ensayo clínico de su nuevo medicamento para la fibrosis pulmonar idiopática. El ensayo no mostró ningún beneficio generalizado, pero sí una mortalidad significativamente menor en el pequeño subconjunto de pacientes con una enfermedad de leve a moderada. Harkonen publicó una nota de prensa con este resultado, y según él este estudio generaría unas ventas enormes. Aunque no dijo nada que fuese objetivamente falso, en 2009 un jurado le condenó por fraude electrónico, con la intención específica de defraudar a los inversores. El Gobierno pedía una sentencia de diez años de cárcel y una multa de veinte mil dólares, pero fue sentenciado a seis meses de arresto domiciliario y tres años de libertad condicional. Un ensayo clínico subsiguiente no descubrió ningún beneficio del uso del medicamento en este subconjunto de

pacientes.[172]

La mala práctica estadística puede ser o no una decisión consciente. Se ha usado deliberadamente para mostrar las limitaciones del proceso científico de publicación y revisión por pares. Johannes Bohannon, del Instituto Alemán de Nutrición y Salud, condujo un estudio en el que los participantes eran asignados aleatoriamente a tres grupos y se le proporcionaba a cada uno una dieta estándar, una dieta baja en carbohidratos o una baja en carbohidratos pero con extra de chocolate. El estudio empleaba toda una batería de indicadores tomados a lo largo de tres semanas. El estudio concluyó que la pérdida de peso en el grupo del chocolate excedió a la del grupo de la dieta baja en carbohidratos en un 10 % (P = 0,04). Este resultado «significativo» se envió a una revista que lo consideró un «manuscrito excepcional» y sugirió que por seiscientos euros «podría ser publicado en nuestra revista principal». La publicación fue recogida en numerosos medios, con titulares como: «El chocolate acelera la pérdida de peso».

Pero entonces se reveló que todo había sido un fraude deliberado. «Johannes Bohannon» era realmente John Bohannon, un periodista; el Instituto de Nutrición y Salud no existía, y el único elemento real eran los datos, que no habían sido inventados. Pero solo había cinco sujetos por grupo, se habían realizado un gran número de pruebas y solo se había informado de las diferencias significativas.

Los autores de este artículo espurio inmediatamente reconocieron el engaño, pero no todos los fraudes estadísticos se cometen para mostrar las debilidades del proceso de revisión por pares.

Fraude deliberado

La invención deliberada de datos existe, pero se cree que es algo relativamente raro. Una revisión de informes anónimos realizados por los propios autores estimó que el 2 % de los científicos admitían falsificación de datos, y lo cierto es que la Fundación Nacional para la Ciencia de Estados Unidos y la Oficina para la Integridad en la Investigación se han ocupado de un número bastante pequeño de actos deliberadamente deshonestos, aunque los detectados deben de ser una subestimación.[173]

Es de lo más coherente que el fraude estadístico pueda ser detectado a través de la ciencia estadística. Uri Simonsohn, un psicólogo de la Universidad de Pensilvania, examinó estadísticos que describían experimentos supuestamente aleatorizados que deberían mostrar la

típica variación aleatoria, pero que eran o bien inverosímilmente similares o diferentes. Por ejemplo, se dio cuenta de que tres desviaciones típicas estimadas citadas en un informe, que se suponía que provenían de diferentes grupos de 15 individuos, eran todas iguales a 25,11. Simonsohn obtuvo los datos en bruto y mostró por medio de una simulación que la probabilidad de obtener unas desviaciones típicas tan similares era increíblemente pequeña. El investigador responsable del informe más tarde dimitió.[174]

Cyril Burt, un psicólogo británico renombrado por su investigación sobre el componente hereditario del cociente intelectual (CI), fue póstumamente acusado de fraude tras descubrir que los coeficientes de correlación que citaba para el caso del CI de gemelos que habían sido criados aparte apenas cambiaban a lo largo del tiempo a pesar de que se trataba de un grupo sistemáticamente creciente de gemelos: la correlación era de 0,770 en 1943, 0,771 en 1955 y 0,771 en 1966. Se le acusó de fabricar los datos, pero todos sus registros habían sido quemados después de su muerte. Es una cuestión todavía discutida; algunos argumentan que debió de ser un error, dado que habría sido un fraude demasiado obvio.

Todo sería más sencillo si la incompetencia total y la deshonestidad fueran los únicos problemas a los que se enfrenta la estadística, por graves que estas sean. Podríamos educar, comprobar, replicar, hacer públicos los datos para que sean examinados y demás, como veremos en el último capítulo acerca de cómo hacer bien los análisis estadísticos. Pero hay un problema mayor y más sutil, que, según algunos, es uno de los grandes responsables de la crisis de reproductibilidad.

«Prácticas científicas cuestionables»

Incluso aunque los datos no hayan sido inventados, el análisis final sea apropiado y los estadísticos y su anejo valor P sean numéricamente correctos, puede ser difícil saber cómo interpretar los resultados si no sabemos exactamente qué han hecho los investigadores en el proceso de llegar a sus conclusiones.

Hemos visto los problemas que se derivan de que los investigadores informen solo de resultados significativos, pero quizá más importante es el conjunto de decisiones conscientes o inconscientes que podría adoptar el investigador dependiendo de lo que parezcan mostrar los datos. Estos «retoques» podrían incluir decisiones sobre cambios en el diseño del experimento, cuándo dejar de reunir datos, qué datos excluir, qué controles incluir, qué grupos enfatizar, qué resultados

destacar, cómo dividir variables continuas en grupos, cómo manejar los valores perdidos, *etc.* Simonsohn llamó a estas decisiones «los grados de libertad del investigador», mientras que Andrew Gelman se refería, más poéticamente, al «jardín de los senderos que se bifurcan». Es probable que todos estos retoques incrementen la probabilidad de obtener significatividad estadística, y todos se pueden agrupar bajo la etiqueta general de «prácticas científicas cuestionables» o QRP.[175]

Es importante distinguir lo que se conoce como estudios exploratorios y estudios confirmatorios. Los experimentos exploratorios son lo que su nombre indica: investigaciones flexibles diseñadas para estudiar muchas posibilidades y sugerir hipótesis que comprobar más adelante a través de estudios confirmatorios más formales. Cualquier retoque es admisible en los estudios exploratorios, pero los confirmatorios deben llevarse a cabo según un protocolo preespecificado y preferiblemente público. Cada uno de ellos puede usar valores P para resumir la fortaleza de la evidencia empleada para sus conclusiones, pero estos valores P deberían distinguirse claramente e interpretarse de manera muy diferente.

Las actividades que tienen como objetivo crear resultados estadísticamente significativos se conocen como «saqueo de valores P» y, aunque la técnica más obvia es llevar a cabo pruebas múltiples e informar de las más significativas, hay formas mucho más sutiles a través de las cuales los investigadores pueden ejercer sus grados de libertad.

¿Escuchar la canción de los Beatles «When I'm Sixty Four» rejuvenece?

Tal vez usted esté muy seguro sobre cuál es la respuesta correcta a esta pregunta. Eso hace que los esfuerzos de Simonsohn y sus colegas por conseguir un resultado positivo y significativo sean aún más impresionantes, aunque hay que admitir que lo consiguieron a través de unos medios bastante dudosos.[176]

Se asignó aleatoriamente a los estudiantes de grado de la Universidad de Pensilvania que escuchasen o bien «When I'm Sixty-Four», de los Beatles, o «Kalimba» o «Hot Potato», de los Wiggles. Después, se preguntó a los estudiantes cuándo habían nacido, lo mayores que se sentían y una serie de preguntas deliciosamente irrelevantes.[177]

Simonsohn y sus colegas analizaron repetidamente los datos de todas las maneras posibles, y siguieron enrolando participantes hasta que encontraron algún tipo de asociación significativa. Esto sucedió después de incluir 34 sujetos, y aunque no había una relación

significativa entre la edad de los participantes y el disco que habían escuchado, solo con comparar «When I'm Sixty-Four» y «Kalimba», se las arreglaron para obtener $P < 0{,}05$ en una regresión que controlaba por la edad del padre. Naturalmente, solo informaron del resultado significativo sin mencionar primero el enorme número de retoques, manipulaciones e información selectiva que habían llevado a cabo. Estos serían revelados al final del artículo, que se ha convertido en una clásica demostración deliberada de la práctica que se conoce ahora como «HARKing»: inventarse la hipótesis después de conocer los resultados.[178]

¿Cuán extendidas están estas prácticas científicas cuestionables?

En una encuesta a 2.155 psicólogos académicos estadounidenses en 2012,[179] solo el 2 % admitió que había falsificado datos. Pero cuando se les preguntó sobre una lista de prácticas científicas cuestionables:

→ El 35 % dijo que habían informado de un resultado inesperado como si hubiese sido esperado desde el principio.

→ El 58 % dijo que se habían puesto a recopilar más datos después de ver si los resultados eran significativos.

→ El 67 % dijo que no habían informado de todas las respuestas obtenidas.

→ El 94 % reconoció al menos una de las prácticas científicas cuestionables que hemos enumerado.

Generalmente argumentaron que estas prácticas eran defendibles; después de todo, ¿por qué no informar de un resultado interesante, aunque inesperado? De nuevo, el problema surge debido a que se borran las fronteras entre los estudios exploratorios y confirmatorios: muchas de las prácticas, incluido el HARKing, podrían ser admisibles en un estudio exploratorio cuyo objetivo deliberado es desarrollar nuevas ideas para ser comprobadas, pero deberían estar estrictamente prohibidas en estudios que afirman probar algo.

Fallo de comunicación

Sea el trabajo estadístico bueno o no tan bueno, en algún momento debe ser comunicado a las audiencias, ya sean estas colegas o el público en general. Los científicos no son los únicos que informan de hallazgos basados en evidencia estadística. Los Gobiernos, los

políticos, las organizaciones benéficas y otras organizaciones no gubernamentales compiten por nuestra atención, y para ello usan cifras y datos científicos para proporcionar una base aparentemente «objetiva» a sus afirmaciones. La tecnología ha cambiado, lo que permite usar diversas fuentes en Internet y de las redes sociales para comunicar resultados con pocos controles que aseguren un uso fiable de la evidencia.

Figura 12.1. Diagrama simplificado de los flujos de información tradicionales desde las fuentes estadísticas hasta el público. En cada

estadio hay filtros derivados de unas prácticas cuestionables de investigación, de interpretación y de comunicación, como la información selectiva, la falta de contextualización, la exageración de la importancia del resultado, *etc.*

La figura 12.1 proporciona una visión sumamente simplificada del proceso a través del cual nos enteramos de la evidencia estadística. [180] Todo empieza con el creador de los datos, y pasa después a las «autoridades», después a sus departamentos de prensa y comunicación, a los periodistas que escriben las historias y a los editores que añaden los titulares, y finalmente a nosotros como miembros individuales de la sociedad. A lo largo de todo el proceso se pueden producir errores y distorsiones.

¿Qué es lo que aparece en la literatura?

El primer filtro se produce en el momento de la publicación del trabajo estadístico llevado a cabo por los investigadores. Muchos estudios ni siquiera se mandan a publicar, o bien porque los resultados no parecen lo suficientemente «interesantes», o porque no encajan en los objetivos de la organización de investigación: las empresas farmacéuticas han sido acusadas en el pasado de ocultar estudios cuyos resultados no les convenían. Esto hace que se desechen datos valiosos, y que la literatura tenga un sesgo positivo. Desconocemos aquello que no se nos cuenta.

Este sesgo positivo empeora porque hay «descubrimientos» que tienen una mayor probabilidad de ser aceptados para su publicación en las revistas más prominentes, por la poca predisposición a publicar réplicas y, por supuesto, por el hecho de que todas las prácticas de investigación cuestionables que hemos visto pueden llevar a una significatividad estadística exagerada.

La oficina de prensa

En el siguiente estadio, cuando las noticias científicas pasan a las oficinas de prensa para intentar conseguir cobertura en los medios, surgen más problemas potenciales. Ya hemos visto cómo una nota de prensa excesivamente entusiasta referida a un estudio sobre la posición socioeconómica y el riesgo de tumores cerebrales llevó al clásico titular: «Por qué ir a la universidad incrementa el riesgo de tumor cerebral». Esa oficina de prensa no está sola en su tendencia a exagerar: un estudio descubrió que de 462 notas de prensa de universidades del Reino Unido en 2011:

→ Un 40 % contenían consejos exagerados.

→ Un 33 % contenían afirmaciones causales exageradas.

→ Un 36 % contenían inferencias exageradas a humanos a partir de investigación en animales.

→ La mayoría de las exageraciones aparecidas en la prensa podían atribuirse a la nota de prensa.

El mismo grupo de investigadores encontró resultados ligeramente más alentadores referidos a las más importantes revistas biomédicas: las afirmaciones causales o las recomendaciones de un artículo eran exageradas en el 21 % de las correspondientes notas de prensa, aunque estas exageraciones, que normalmente eran denunciadas, no generaron más cobertura en la prensa.[181]

Vimos en el capítulo 1 cómo los «marcos» que aplicamos a las cifras pueden influir en su interpretación: por ejemplo, «90 % sin grasa» suena algo mejor que «10 % de grasa». Un buen ejemplo de cómo echarle imaginación a la forma de contar historias se produjo cuando un estudio valioso pero bastante aburrido descubrió que el 10 % de las personas son portadoras de un gen que las protege contra la hipertensión. El equipo de comunicación reformuló este resultado como que «nueve de cada diez personas llevan un gen que incrementa el riesgo de hipertensión»: este mensaje presentado negativamente recibió la pertinente cobertura de prensa internacional.[182]

Los medios

Se suele culpar a los periodistas por una mala cobertura de las historias científicas y estadísticas, pero en gran medida están a merced de lo que se les proporciona en las notas de prensa y los artículos científicos, y en cómo los titulares escogidos por los editores enmarcan sus historias. Pocos lectores de periódicos se dan cuenta de que la persona que escribió el artículo generalmente tiene un control mínimo sobre el titular, y, por supuesto, los titulares son los que atraen lectores.

El principal problema de la cobertura de prensa no son las falsedades totales y absolutas, sino la manipulación y la exageración a través de una inapropiada interpretación de los «hechos»: estos pueden ser técnicamente correctos, pero están distorsionados por lo que podríamos llamar «unas prácticas de interpretación y comunicación cuestionables». A continuación enumeraré una lista corta de las formas a través de las cuales los medios condimentan su cobertura de historias con contenido estadístico. Muchas de estas prácticas cuestionables serían consideradas defendibles por aquellos cuya carrera profesional depende de atraer lectores, oyentes o clics.

1. Escoger historias que van en contra del consenso actual.

2. Promover historias sin tener en cuenta su calidad.

3. No informar sobre incertidumbres.

4. No proporcionar contexto o perspectiva comparada, como, por ejemplo, la tendencia a largo plazo.

5. Sugerir una causa cuando solo se observa una asociación.

6. Exagerar la relevancia e importancia de los descubrimientos.

7. Afirmar que la evidencia da apoyo a una determinada política.

8. Usar marcos positivos o negativos dependiendo de si el objetivo es tranquilizar o asustar.

9. No tener en cuenta los conflictos de interés o puntos de vista alternativos.

10. Usar un gráfico vistoso pero no informativo.

11. Informar solo de los riesgos relativos, no de los absolutos.

Esta última práctica es universal. Vimos en el capítulo 1 cómo podía conseguirse que una historia sobre cómo el beicon incrementa el riesgo de cáncer colorrectal sonase más impresionante citando los riesgos relativos en lugar de los absolutos. Los periodistas saben que los riesgos relativos, a los que los medios se refieren simplemente como un «aumento del riesgo» con independencia de su magnitud, son una forma efectiva de hacer que una historia parezca más excitante, y no ayuda el hecho de que los riesgos relativos en forma de razones de probabilidad, razones de tasas de incidencia y razones de riesgo sean el resultado estándar para la mayor parte de los estudios biomédicos. El impactante titular «Por qué un atracón de televisión puede matarle» se derivaba de un estudio epidemiológico que estimaba un riesgo relativo ajustado de 2,5 de embolia pulmonar crítica asociada con ver más de 5 horas de televisión cada noche comparado con menos de 2,5 horas. Pero un cuidadoso escrutinio de la tasa absoluta en el grupo de riesgo alto reveló que podías ver más de 5 horas de televisión cada noche durante 12.000 años antes de experimentar el evento. Esto, de alguna manera, reduce algo su impacto.[183]

Este titular se escribió para atraer la atención y conseguir clics, y consiguió su objetivo —yo, ciertamente, lo encontré irresistible—.

Cuando buscamos novedades y estimulación inmediata, no es sorprendente que los medios sazonen algo los estudios científicos y favorezcan las afirmaciones inusuales (y probablemente exageradas) sobre la sólida evidencia estadística.[184] En el siguiente capítulo veremos cómo se puede mejorar todo esto, pero primero volvamos a las sorprendentes afirmaciones de Daryl Bem sobre la precognición.

Daryl Bem sabía que estaba publicando unos resultados extraordinarios, y, en un gesto que le honra, fomentó la réplica de su estudio y proporcionó los materiales necesarios para ello. No obstante, cuando otros investigadores le tomaron la palabra, e intentaron (y no consiguieron) reproducir sus resultados, la revista que había publicado el estudio original de Bem se negó a publicar las réplicas fallidas.

Entonces ¿cómo llegó Bem a sus resultados? En distintos momentos ajustó su diseño en respuesta a los datos, y escogió destacar unos grupos particulares, por ejemplo, informar de la precognición positiva cuando mostraba fotos eróticas y no los resultados negativos de las fotos no eróticas. Bem reconoció que «empezaba uno [un experimento], y si no iba a ningún lado, lo abandonaba y lo reiniciaba con cambios». En el artículo se informaba de algunos de estos cambios, pero no de otros.[185] Andrew Gelman observó que:

las conclusiones [de Bem] se basan en valores P, que son afirmaciones referidas a cómo serían los resúmenes de datos si los datos hubiesen salido de otra manera, pero Bem no ofrece ninguna evidencia de que si los datos hubiesen salido de otra manera, sus análisis hubiesen sido los mismos. De hecho, los nueve estudios de su artículo incluyen todo tipo de análisis de datos distintos.[186]

Su caso es un ejemplo clásico de alguien que está explotando demasiados grados científicos de libertad. Pero Bem prestó un gran servicio a la psicología y a la ciencia en general: su artículo de 2011 fue el catalizador de un examen de conciencia colectivo por parte de los científicos sobre las posibles razones de la falta de fiabilidad de la literatura científica. Se sugirió incluso que todo el ejercicio, como otros estudios mencionados en este capítulo, fue deliberadamente planeado por Bem para revelar las debilidades de la investigación en psicología.

Resumen

→ La mala práctica estadística tiene alguna responsabilidad por la crisis de reproductibilidad en la ciencia.

→ La invención deliberada de datos parece ser bastante rara, pero los errores en los métodos estadísticos son frecuentes.

→ Un problema aún mayor son las prácticas cuestionables de investigación que tienden a llevar a afirmaciones exageradas de significatividad estadística.

→ En el proceso a través del cual la evidencia estadística llega al público, las oficinas de prensa, los periodistas y los editores empeoran aún más las afirmaciones estadísticas injustificadas a través de su uso de interpretaciones y prácticas comunicativas cuestionables.

[165] Open Science Collaboration, «Estimating the Reproductibility of Psychological Science», Science 349: 6251 (28 de agosto de 2015), aac4716.

[166] A. Gelman y H. Stern, «The Difference Between "Significant" and "Not Significant" Is Not Itself Statistically Significant», American Statistician 60:4 (noviembre de 2006), 328-31.

[167] La caída comenzó poco después de la creación de Facebook, pero los datos no nos dicen si esto es simple correlación o causación.

[168] Ronald Fisher, conferencia presidencial para el primer Congreso Estadístico Indio, 1938, Sankhya 4 (1938), pp. 14-17.

[169] Se ha argumentado que este error, combinado con otras críticas, cambia los resultados del estudio, pero esto es firmemente rechazado por los autores originales. Véase «The Reinhart and Rogoff Controversy: A Summing Up», New Yorker, 26 de abril de 2013.

[170] Por sus siglas en inglés, Securities and Exchange Commission. (N. del T.).

[171] «AXA Rosenberg encuentra un error de codificación en un programa de riesgos», Reuters, 24 de abril de 2010.

[172] La historia de Harkonen se recoge en «The Press-Release Conviction of a Biotech CEO and its Impact on Scientific Research», Washington Post, 13 de septiembre de 2013.

[173] D. Fanelli, «How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data», PLOS ONE 4:5 (29 de mayo de 2009), e5738.

[174] U. Simonsohn, «Just Post It: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone», Psychological Science 24:10 (octubre de 2011), pp. 1875-1888.

[175] Por sus siglas en inglés, Questionable Research Practices. (N. del T.).

[176] J. P. Simmons, L. D. Nelson y U. Simonsohn, «False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant», Psychlogical Science 22:11 (noviembre de 2011), pp. 1359-1366.

[177] Entre otras, cuánto les gustaría comer en una cafetería, la raíz cuadrada de 100, si estaban de acuerdo con la idea de que «los ordenadores son máquinas complicadas», la edad de su padre, la edad de su madre, si se aprovecharían de una oferta especial, su orientación política, quién de cuatro mariscales de campo canadienses piensan que ganó un premio, con qué frecuencia se referían al pasado como «qué tiempos aquellos», y así sucesivamente.

[178] «HARK» por sus siglas en inglés, «Hypothesis After the Results are Known». (N. del T.).

[179] L. K. John, G. Loewenstein y D. Prelec, «Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling», Psychological Science 23:5 (mayo de 2012), pp. 524-532.

[180] D. Spiegelhalter, «Trust in Numbers», Journal of the Royal Statistical Society: Series A (Statistics in Society) 180: 4 (2017), pp. 948-965.

[181] P. Sumner et al., «The Association Between Exaggeration in Health Related Science News and Academic Press Releases: Retrospective Observational Study», British Medical Journal 349 (10 de diciembre de 2014), g7015.

[182] «Nine in 10 People Carry Gene Which Increases Chance of High Blood Pressure», Daily Telegraph, 15 de febrero de 2010.

[183] «Why Binge Watching Your TV Box-Sets Could Kill You», Daily Telegraph, 25 de julio de 2016.

[184] Yo a veces sigo el que podría llamarse «principio de Groucho», por la paradójica afirmación de Groucho Marx de que nunca se uniría a un club que aceptase a gente como él. Debido a que las historias han

pasado por tantos filtros que favorecen la distorsión y la selección, el hecho mismo de que escuche una afirmación basada en estadísticas es razón suficiente para no creérmela.

[185] En otro artículo se recoge otra cita de Bem: «Estoy completamente a favor del rigor [...] pero prefiero que otras personas se encarguen de él. Veo su importancia —y para algunas personas es divertido— pero no tengo paciencia para ello [...] Si miras todos mis experimentos pasados, siempre eran estrategias retóricas. Reunía datos para mostrar cómo podían demostrarse mis argumentos. Usaba los datos como un arma de persuasión, y realmente nunca me preocupé sobre si podían replicarse o no». Esta cita de Bem aparece en «Daryl Bem Proved ESP Is Real: Which Means Science is Broken», Slate, 17 de mayo de 2017.

[186] El sucinto resumen de Gelman fue que «el estudio de Bem es una basura».

13

Cómo podemos hacer mejor los análisis estadísticos

¿Cuál es el beneficio de las revisiones de cáncer de ovarios?

En 2015 se publicaron los resultados de un enorme examen de cáncer de ovarios realizado en el Reino Unido. Había comenzado en 2001, cuando, después de un cuidadoso cálculo de la potencia necesaria, alrededor de doscientas mil mujeres fueron asignadas aleatoriamente a uno de dos modos de exploración del cáncer de ovarios, o a un grupo de control. Los investigadores preespecificaron de manera rigurosa un protocolo en el cual el análisis primario era una reducción de la mortalidad por cáncer de ovarios, evaluado usando un método estadístico que asumía que la reducción proporcional del riesgo sería la misma durante todo el período de seguimiento.[187]

Cuando los datos fueron analizados finalmente, después de un seguimiento medio de once años, el análisis primario preespecificado no mostró un beneficio estadísticamente significativo, y los autores informaron debidamente de este resultado no significativo como su principal conclusión. Entonces, ¿por qué el periódico Independent publicó el titular «Avance en las pruebas sanguíneas para cáncer de ovarios: el enorme éxito del nuevo método de prueba podría llevar a una revisión nacional en Gran Bretaña»?[188]

Volveremos más adelante a si los resultados de este estudio masivo, y muy caro, fueron interpretados adecuadamente.

En el capítulo anterior vimos cómo a lo largo de todo el proceso que siguen los estudios estadísticos se pueden producir malas prácticas, lo que significa que si queremos que el uso de la estadística mejore, hay tres grupos que tienen que actuar:

→ Productores de estadísticas: como los científicos, los estadísticos, las compañías demoscópicas y la industria. Pueden hacer mejor sus estadísticas.

→ Comunicadores: como las revistas científicas, las organizaciones benéficas, los departamentos gubernamentales, las oficinas de prensa, los periodistas y los editores. Pueden comunicar mejor la estadística.

→ Audiencias: como el público en general, los diseñadores de las políticas y los profesionales. Pueden evaluar mejor las estadísticas.

Vamos a ver con más detalle qué puede hacer cada uno de estos grupos.

Mejorar lo que se produce

¿Cómo podemos conseguir que todo el proceso científico sea más digno de confianza? Varios eminentes investigadores han colaborado para generar un «manifiesto de la reproductibilidad», que incluye una mejora de los métodos y de la formación investigadora; la promoción del prerregistro del diseño y el análisis de los estudios; una mejor comunicación de lo que se está haciendo; y el favorecimiento de la réplica, de una revisión por pares diversificada y de la apertura y la transparencia.[189] Muchas de estas ideas se reflejan en la Estructura para la Ciencia Abierta, una infraestructura que favorece particularmente el intercambio de datos y el prerregistro de los estudios.[190]

Si se tienen en cuenta los ejemplos presentados en el capítulo anterior, no debería ser una sorpresa que muchas de las sugerencias del manifiesto se refieran a la práctica estadística. En particular su llamamiento a prerregistrar los estudios tiene como objetivo prevenir el tipo de comportamiento que se ha ilustrado tan claramente en el anterior capítulo, por el cual el diseño, las hipótesis y los análisis del estudio se adaptan a los datos a medida que estos van llegando. Pero también se podría argumentar que una preespecificación completa no es realista y va en contra de la imaginación del investigador, y que la flexibilidad para adaptarse a nuevos datos puede ser una característica positiva. Una vez más, la respuesta está en una distinción clara entre los estudios exploratorios y confirmatorios, siempre informando claramente de la secuencia de decisiones que han tomado los investigadores.

La preespecificación de un análisis no está exenta de problemas, dado que puede constreñir a los investigadores a un análisis que, cuando llegan los datos, resulta ser inapropiado. Por ejemplo, el equipo que llevó a cabo el estudio sobre prevención del cáncer de mama planeó incluir todos los pacientes aleatorizados en su análisis, pero descubrió que si los casos «prevalentes» (los que se descubrió que tenían cáncer de ovarios antes de que comenzase el estudio) eran excluidos del análisis, lo que podría ser considerado algo bastante razonable, entonces la estrategia de revisión multimodal sí que mostraba una reducción significativa del 20 % de la mortalidad por cáncer de ovarios ($P = 0{,}02$). Además, incluso cuando se incluyen todos los casos con independencia de si tenían o no cáncer de ovarios al comienzo del estudio, también apareció una reducción del 23 % en

mortalidad en el grupo multimodal en el período entre 7 y 14 años después de la aleatorización. De manera que cuestiones que podían no haber sido previstas, como que se hubiese incluido en la aleatorización a personas que ya tenían cáncer de ovarios, o que la revisión tarde algo en ser efectiva, impidieron que el resultado preplaneado fuese significativo.

Los autores fueron meticulosos en informar de que su análisis primario no mostraba un resultado significativo, y se lamentaron de que «la principal limitación de este estudio fue nuestra incapacidad de anticipar el efecto tardío de la revisión en nuestro diseño estadístico». Esto no impidió que parte de los medios interpretasen un resultado no significativo como una confirmación de la hipótesis nula, informando de que la revisión de cáncer de ovarios no tenía ningún efecto. El titular del Independent según el cual esa misma revisión podía salvar miles de vidas, aunque un tanto osada, quizá reflejaba mejor las conclusiones del estudio.

Mejorar la comunicación

Este libro ha presentado ejemplos de coberturas informativas bastante calamitosas de historias basadas en estadísticas. No hay una forma sencilla de influir en el periodismo y los medios de comunicación, particularmente en una época en la que la industria se ve desafiada por la competición de las redes sociales y las publicaciones digitales no reguladas, y los ingresos por publicidad están disminuyendo. No obstante, los estadísticos han colaborado en la elaboración de guías maestras para los medios de comunicación y en programas de formación para periodistas y oficinas de prensa. Las buenas noticias son que el periodismo basado en datos está en auge, y la colaboración con los periodistas puede llevar a historias más ricas basadas en datos, con narrativas y visualizaciones apropiadas y atractivas.

No obstante, hay riesgos inherentes en la conversión de números en historias. Los reportajes tradicionales necesitan un contenido emocional, una trama bien hilada y unas conclusiones bien delineadas; la ciencia raramente produce esas cosas. Por ello, la simplificación y las exageraciones son tentadoras. Las historias deben reflejar la evidencia: sus fortalezas, debilidades e incertidumbres. Idealmente las historias podrían contar que un medicamento u otra intervención médica no es ni buena ni mala: tiene beneficios y efectos secundarios, que la gente puede sopesar de maneras distintas, y, razonablemente, llegar a conclusiones distintas. Los periodistas parecen huir de ese tipo de narrativas sutiles, pero lo cierto es que un buen comunicador debería ser capaz de hacer que esas historias

fueran atractivas (por ejemplo, incluyendo testimonios de personas con puntos de vista distintos). En este sentido, Christie Aschwanden, de FiveThirtyEight, cuando discutía las estadísticas sobre revisiones por cáncer de mama afirmó que había decidido no someterse a dicho procedimiento, pero también comentó que una amiga suya muy inteligente, enfrentada a exactamente la misma evidencia, había tomado la decisión opuesta.[191] Este enfoque pone de relieve la importancia de los valores e inquietudes personales, pero respetando al mismo tiempo la evidencia estadística.

También podríamos esforzarnos más en investigar cómo mejorar la comunicación de la estadística. Por ejemplo, cómo podemos comunicar mejor la incertidumbre sobre hechos y sobre el futuro sin poner en cuestión la confianza y la credibilidad, y cómo podemos adaptar nuestras técnicas a audiencias con diferentes actitudes y conocimiento. Estas cuestiones son importantes y pueden ser objeto de investigación. Además, el calamitoso nivel del debate estadístico en la campaña del referéndum sobre el Brexit en el Reino Unido sugiere que hay que investigar cuáles son las distintas formas a través de las cuales comunicar cómo pueden impactar en la sociedad las decisiones políticas.

Ayudar a denunciar las malas prácticas

Muchas personas y grupos juegan un papel en la identificación de las malas prácticas estadísticas: los evaluadores de artículos enviados para ser publicados, los que llevan a cabo revisiones sistemáticas de evidencia publicada, los periodistas y las organizaciones de comprobación de la veracidad de las noticias y las personas particulares.

Uri Simonsohn ha sido particularmente claro al sugerir que los revisores deben ser más estrictos y asegurarse de que los autores de un artículo han cumplido con los requisitos establecidos por la revista, que puedan demostrar que sus resultados son firmes y no dependen de decisiones arbitrarias. Si aun así queda alguna duda, los revisores deberían exigir réplicas de los análisis. Por otro lado, sugiere también que los revisores deberían ser más tolerantes ante imperfecciones en los resultados, algo que, a su vez, debería favorecer que estos sean comunicados honestamente.[192]

Pero, desde la experiencia de alguien que ha revisado personalmente cientos de artículos científicos, he de decir que no siempre es fácil identificar los problemas. Las guías proporcionadas por las revistas pueden ser útiles, pero también pueden ser manipuladas por los

autores para hacer que el artículo parezca razonable. Tengo que admitir que he desarrollado una «nariz» fina, que olfatea en búsqueda de indicios de que, por ejemplo, solo se informe de las comparaciones «interesantes» de entre las muchas que se han realizado.

Una nariz así reaccionará sin duda a la vista de unos resultados demasiado buenos para ser verdad, como por ejemplo si se observa un gran efecto con una muestra pequeña. Un ejemplo típico es un estudio de 2007 ampliamente difundido que afirmaba demostrar que las personas atractivas tienen más hijas. El «atractivo físico» era evaluado en una escala de cinco puntos en una encuesta estadounidense a adolescentes, y quince años más tarde, resultaba que aquellos considerados como «muy atractivos» solo tenían un 44 % de primogénitos varones, frente al 52 % estándar de la gente más normalita (como mostró Arbuthnot, de media nacen ligeramente más niños que niñas). Este descubrimiento es estadísticamente significativo, pero tal como percibió Andrew Gelman, es un efecto demasiado grande como para ser plausible, y solo se da en el grupo «más atractivo». No hay nada en el artículo que vaya a revelar la total implausibilidad de este resultado: se requiere conocimiento externo. [193]

Sesgos de publicación

Los científicos examinan un número enorme de artículos publicados cuando llevan a cabo revisiones sistemáticas, con el fin de reunir toda la literatura y sintetizar el estado actual del conocimiento. Esa empresa resultará irremediablemente fallida si lo que se ha publicado es un subconjunto sesgado de los trabajos que se han llevado a cabo, por ejemplo porque los resultados negativos no hayan sido enviados a publicar, o porque unas prácticas científicas cuestionables hayan llevado a un exceso no justificado de resultados significativos.

Se han desarrollado técnicas estadísticas para identificar ese sesgo de publicación. Supongamos que tenemos un conjunto de estudios cuyo objetivo es comprobar la misma hipótesis nula, que afirma que una intervención no tiene efectos. Con independencia de los experimentos que realmente fueron llevados a cabo, si la intervención no tiene efectos, puede probarse teóricamente que cualquier valor P que compruebe la hipótesis nula es igualmente probable que adopte cualquier valor entre 0 y 1, de manera que los valores P de muchos estudios que comprueben ese efecto deben tender a estar distribuidos uniformemente. Por el contrario, si realmente hay un efecto, los valores P tenderán a estar sesgados hacia valores pequeños.

La «curva P» supone observar los valores P reales presentados en las pruebas de significatividad; es decir, cuando P < 0,05. Hay dos hechos que pueden levantar sospechas. En primer lugar, si hay un grupo de valores P que están justo por debajo de 0,05, lo que sugiere es que se han «cocinado» un poco los datos para que algunos pasen esta frontera crucial. En segundo lugar, supóngase que estos valores P significativos no se concentran en dirección al 0, sino que están distribuidos uniformemente entre 0 y 0,05. Esta sería justo la pauta que se daría si la hipótesis nula fuese cierta, mientras que los únicos resultados de los que se informa como significativos son apenas 1 de cada 20, aquellos que, por pura suerte, hacen que el valor P se deslice por debajo de 0,05. Simonsohn y otros echaron un vistazo a la literatura publicada en psicología que apoyaba la idea popular de que dar a la gente excesivas opciones llevaba a consecuencias negativas. Un análisis de la curva P sugería que había un sustancial sesgo de publicación y que no existía una clara evidencia a favor de ese efecto. [194]

La evaluación de una afirmación o de una historia basada en la estadística

Ya seamos periodistas, encargados de comprobar la veracidad de las noticias, académicos, profesionales de organismos públicos, de la empresa privada o de organizaciones no gubernamentales, o simplemente personas particulares, somos destinatarios habituales de afirmaciones basadas en evidencia estadística. Saber si estas afirmaciones son ciertas o no es una destreza vital en el mundo moderno.

Partamos del osado supuesto de que todos aquellos implicados en la recopilación, análisis y uso de la estadística suscriben un marco ético en el cual es fundamental la confianza. Onora O'Neill, una eminente filósofa kantiana y una autoridad en el campo de la confianza, ha señalado que las personas no deberían intentar que se confiase en ellas, dado que esto te lo dan otros, sino más bien demostrar lo digno de confianza que es su trabajo. O'Neill ha proporcionado una lista corta y de lo más perspicaz de los requisitos necesarios para ello: por ejemplo, la fiabilidad requiere honestidad, competencia y seriedad. Pero también requiere, en su opinión, evidencia de que el trabajo realizado es digno de confianza, lo que significa que es transparente, lo cual no se consigue simplemente ahogando a la gente en un mar de datos, sino más bien siendo «inteligentemente transparente».[195] Esto significa que las afirmaciones basadas en datos tienen que ser:

→ Accesibles: el público debe poder acceder a la información.

→ Inteligibles: el público debe poder comprender la información.

→ Evaluables: si así lo desea, el público debe poder comprobar la fiabilidad de las afirmaciones.

→ Usables: el público debe poder explotar la información para sus necesidades.

Pero evaluar lo digna de confianza que es una afirmación no es una tarea sencilla. Los estadísticos, y no solo ellos, se han pasado décadas aprendiendo cómo sopesar sus afirmaciones, y cómo plantear preguntas que ayuden a identificar fallos. No es cuestión de una simple lista de tareas: se requiere experiencia y una cierta actitud escéptica. Teniendo en cuenta esto, presento aquí una serie de preguntas que intentan recoger toda la sabiduría que pueda contener este libro. Los términos y las cuestiones referidas a cada una de ellas son o bien obvios, o bien ya se han analizado. A mí esta lista me resulta útil, espero que a usted también.

Diez preguntas que hay que hacer cuando nos enfrentamos a una afirmación basada en evidencia estadística

¿Son las cifras dignas de confianza?

1. ¿Es el estudio riguroso? Por ejemplo, comprobar la «validez interna», que el diseño y la redacción de las preguntas son los apropiados, que se haya prerregistrado el protocolo, que se haya tomado una muestra representativa mediante la aleatorización, y que se haya realizado una comparación justa con un grupo de control.

2. ¿Cuál es la incertidumbre estadística o la confianza sobre los resultados? Compruebe los márgenes de error, los intervalos de confianza, la significatividad estadística, el tamaño de la muestra, las comparaciones múltiples, los sesgos sistemáticos.

3. ¿Es apropiado el resumen? Compruebe el uso apropiado de las medias, la variabilidad, los riesgos relativo y absoluto.

¿Es la fuente digna de confianza?

4. ¿Es fiable la fuente? Considérese la posibilidad de que la fuente esté sesgada, con conflictos de interés, y compruébese que la publicación cuenta con revisión por pares independiente. Pregúntese por qué esta fuente quiere que usted escuche esta historia.

5. ¿Está la historia siendo manipulada? Tenga en cuenta el uso de

marcos, apelaciones emocionales mediante anécdotas sobre casos extremos, gráficos que llevan a confusión, titulares exagerados, números que suenan demasiado bien.

6. ¿Qué es lo que no se me está contando? Esta es quizás la pregunta más importante de todas. Considere la posibilidad de que los resultados hayan sido seleccionados a conveniencia de los autores, que se haya excluido información que vaya en contra de su historia y que no haya comentarios independientes.

¿Es la interpretación digna de confianza?

7. ¿Cómo encaja la afirmación con lo que ya conocemos? Considérese el contexto, si las comparaciones son apropiadas, incluyendo los datos históricos y qué han mostrado los otros estudios, idealmente mediante un metaanálisis.

8. ¿Cuál es la explicación de los autores de los resultados? En este caso hay que fijarse en cuestiones como la correlación versus causalidad, la regresión a la media, la afirmación inapropiada de que un resultado no significativo significa que «no hay efecto», las variables de confusión, la atribución, la falacia del fiscal.

9. ¿Cuál es la relevancia de la historia para el público? Piénsese en la generalización de los resultados, si las personas objeto de estudio son un caso especial, si ha habido una extrapolación de ratones a hombres.

10. ¿Es el efecto predicado importante? Compruébese si la magnitud del efecto es prácticamente significativa, y tenga especial cuidado ante afirmaciones de «incremento del riesgo».

Ética aplicada a los datos

La creciente preocupación sobre el mal uso de los datos personales, particularmente cuando estos se toman de cuentas en redes sociales, ha hecho que cobren importancia los aspectos éticos de la ciencia de los datos y la estadística. Mientras que los expertos en estadística de organismos públicos están sujetos a un código de conducta reglado, la disciplina más general de la ética aplicada a los datos está todavía en sus estadios iniciales.

Este libro se ha ocupado de la necesidad de que los algoritmos que afectan a la gente sean justos y transparentes, de la importancia de hacer ciencia honesta y reproducible, y de los requisitos de una comunicación digna de confianza. Todo esto forma parte de la ética de

los datos. Las historias que hemos contado muestran el daño que provocan los conflictos de intereses, o incluso el simple exceso de entusiasmo. Se podrían haber incluido muchos otros temas: la privacidad y la propiedad de los datos, el consentimiento informado para su uso más amplio, los aspectos legales de la explicación de algoritmos, *etc*.

Aunque la ciencia estadística pueda parecer un tema sumamente técnico, siempre tiene lugar en el contexto de una sociedad ante la cual sus exponentes tienen una responsabilidad. En el futuro cercano podemos esperar que la ética de los datos formará una parte integral de la formación en estadística.

Un ejemplo de buena ciencia estadística

Antes de las elecciones generales del Reino Unido del 8 de junio de 2017, la mayoría de las encuestas sugerían que los conservadores obtendrían una mayoría sustancial. Minutos después de que finalizase la votación a las 22.00 h, un equipo de estadísticos predijo que los conservadores habían perdido muchos escaños y la mayoría, lo que significaba que habría un parlamento sin mayoría absoluta. Esta afirmación fue recibida con incredulidad. ¿Cómo hizo este equipo una predicción tan osada? ¿Tenía razón?

Parece de lo más apropiado que un libro que ha intentado celebrar las buenas prácticas en el arte y la ciencia de los datos, y no simplemente regodearse en estudios erróneos, termine con un gran ejemplo de ciencia estadística.

Podría parecer un tanto curioso preguntarse quién ganó las elecciones inmediatamente después de que estas hayan tenido lugar: después de todo, podemos simplemente esperar sentados toda la noche a que salgan los resultados. Pero se ha convertido en parte del espectáculo de las elecciones el que a los pocos minutos de que cierren las urnas, los expertos estén haciendo predicciones sobre cuál será el resultado. Nótese que los resultados ya son invariables, pero en ese momento son desconocidos, por lo que nos encontramos ante un ejemplo típico del tipo de incertidumbre epistémica que surge cuando hablamos de las tasas de desempleo y otras cifras que están «ahí fuera», pero desconocemos.

Consideremos el ciclo PPDAC. El Problema es predecir rápidamente los resultados de las elecciones del Reino Unido a los pocos minutos de que se cierren las urnas. El equipo, compuesto por los estadísticos David Firth y Jouni Kuha y el experto en elecciones John Curtice,

diseñó un Plan para llevar a cabo encuestas a pie de urna, en las cuales alrededor de 200 votantes eran entrevistados al abandonar 144 de los alrededor de 40.000 colegios electorales —los mismos 144 colegios electorales, y esto es lo importante, que se habían usado en encuestas a pie de urna anteriores—. Los Datos incluían las respuestas de los participantes sobre a quién habían votado, y, más importante aún, a quién habían votado en las elecciones anteriores.

El Análisis usa un repertorio de técnicas que podemos ver mejor a través de los estadios del proceso de derivación de inferencias establecidos en el capítulo 3.

→ De los datos a la muestra. La experiencia sugiere que las respuestas deberían ser una medida razonablemente fiable de lo que la gente ha votado en esta y en las elecciones anteriores, dado que se trata de encuestas a pie de urna y los entrevistados dicen lo que han hecho y no lo que van a hacer.

→ De la muestra a la población objeto de estudio. Se toma una muestra representativa a aquellos que han votado realmente en cada colegio electoral, de manera que los resultados de la muestra se pueden usar para estimar aproximadamente el cambio en el voto en esa pequeña área.

→ De la población objeto de estudio a la población objetivo. Se usa el conocimiento de los datos demográficos de cada colegio electoral, se construye un modelo de regresión que intenta explicar cómo la proporción de gente que ha cambiado su voto entre elecciones depende de las características de los votantes en esa zona electoral. En este sentido, no tiene que asumirse que el cambio en el voto sea el mismo en todo el país, sino que puede variar de una zona a otra, por ejemplo, en función de si es una zona rural o urbana. Después, mediante el modelo de regresión estimado, el conocimiento de los datos demográficos de cada una de las aproximadamente 600 circunscripciones y la votación en las elecciones anteriores, se puede hacer una predicción de la votación en estas elecciones para cada circunscripción individual, aunque no se haya entrevistado realmente a ningún votante en la mayoría de las circunscripciones. Este es en esencia el procedimiento de regresión multinivel y post-estratificación que vimos en el capítulo 11.

Lo limitado de la muestra significa que hay incertidumbre en torno a los coeficientes en el modelo de regresión, lo que, cuando se traslada a toda la población de votantes, genera distribuciones de probabilidad de cómo ha votado la gente, y, por tanto, la probabilidad de que cada

candidato obtenga el número máximo de votos. Si sumamos todo esto a lo largo de todas las circunscripciones, nos da el número esperado de escaños, cada uno de los cuales tiene una incertidumbre asociada (aunque los márgenes de error no fueron comunicados en la noche de las elecciones).[196]

La tabla 13.1 muestra las predicciones y los resultados finales de las elecciones de junio de 2017. El número de escaños predicho está notablemente cercano, con un máximo de cuatro escaños de diferencia, al resultado real para cada partido. La tabla muestra que en las tres últimas elecciones en el Reino Unido, esta sofisticada metodología estadística ha sido excepcionalmente precisa. En 2015 predijo enormes pérdidas para los liberal-demócratas: estimó una caída de 57 escaños a tan solo 10, y el prominente político liberal demócrata Paddy Ashdown dijo en una entrevista en directo en televisión que «se comería su sombrero» si esas estimaciones eran correctas. En realidad, los liberal-demócratas sacaron solo 8 escaños.[197]

Año	Escaños	Conservadores	Laboristas	Liberal-demócratas	Nacionalistas escoceses	Otros
2010	Predichos	307	255	59		29
	Reales	307	258	57		28
2015	Predichos	316	239	10	58	27
	Reales	331	232	8	56	23
2017	Predichos	314	266	14	34	21
	Reales	318	262	12	35	22

Tabla 13.1. Predicciones de la encuesta a pie de urna del número de escaños conseguidos por cada partido en las tres últimas elecciones nacionales del Reino Unido al cierre de las votaciones, comparadas con los resultados reales observados. Las predicciones son estimaciones y tienen márgenes de error asociados.

Los medios de comunicación solo proporcionaban una estimación del número de escaños predicho, pero con un margen de error de 20 escaños. En el pasado habían sido más precisos, por lo que se podría decir que el equipo de estadísticos fue bastante afortunado. Pero se merecen esa suerte: han demostrado maravillosamente cómo la ciencia estadística puede llevar a rotundas conclusiones con capacidad de impresionar tanto al público como a los profesionales. Estos destinatarios tienen poca idea de la complejidad de los métodos subyacentes, y del hecho de que estos resultados tan extraordinarios se deben a la atención meticulosa al detalle a través de todo el ciclo de

resolución de problemas.

Resumen

→ Los productores, comunicadores y público desempeñan un papel en la mejora de cómo se usa en la sociedad la ciencia estadística.

→ Los productores necesitan asegurarse de que la ciencia es reproducible. Para demostrar que es digna de confianza, la información debería ser accesible, inteligible, evaluable y usable.

→ Los comunicadores deben ser cautos en no intentar ajustar el relato estadístico a las narrativas estándar.

→ El público debe cuestionar las malas prácticas preguntándose sobre lo dignas de confianza que son sus cifras, sus fuentes y su interpretación.

→ Cuando nos enfrentamos a una afirmación basada en evidencia estadística, primero tenemos que ver si parece plausible.

[187] I. J. Jacobs et al., «Ovarian Cancer Screening and Mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): A Randomised Controlled Trial», The Lancet 387: 10022 (5 de marzo de 2016), pp. 945-956.

[188] «Ovarian Cancer Blood Tests Breakthrough: Huge Success of New Testing Method Could Lead to National Screening in Britain», Independent, 5 de mayo de 2015.

[189] M. R. Munafò *et al.*, «A Manifesto for Reproducible Science», Nature Human Behavior 1 (2017), aoo21.

[190] Open Science Framework: https://osf.io/.

[191] Esta historia está tomada del programa «Science Won't Settle the Mammogram Debate», FiveThirtyEight, 20 de octubre de 2015.

[192] J. P. Simmons, L. D. Nelson y U. Simonsohn, «False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant», Psychological Science 22:11 (noviembre de 2011), pp. 1359-1366.

[193] A. Gelman y D. Weakliem, «Of Beauty, Sex and Power», American Scientist 97:4 (2009), pp. 310-316.

[194] U. Simonsohn, L. D. Nelson y J. P. Simmons, «P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results», Perspectives on Psychological Science 9:6 (noviembre de 2014), pp. 666-681.

[195] Si quiere saber más sobre transparencia inteligente, véase Royal Society, Science as an Open Enterprise, 2012. La visión de Onora O'Neill sobre qué es ser digno de confianza está explicada brillantemente en su charla TedX «What We Don't Understand About Trust», junio de 2013.

[196] La metodología para las encuestas a pie de urna ha sido explicada por David Firth en: https://warwick.ac.uk/fac/sci/statistics/staff/academic-research/firth/exit-poll-explainer/.

[197] No queda constancia de que Paddy Ashdown cumpliese su promesa, aunque esta todavía genera burlas a su costa. En una tertulia radiofónica en la que yo estaba debatiendo esta encuesta, nos trajeron un gran sombrero de chocolate para compartir.

14

En conclusión

Por decirlo claramente, la estadística puede ser difícil. Aunque he intentado referirme a cuestiones de fondo en este libro, y he procurado no enredarme en detalles técnicos, he tenido que basarme inevitablemente en algunos conceptos complejos. De manera que felicidades por haber llegado al final.

En lugar de intentar reducir los capítulos anteriores a una lista corta de sabios consejos, puedo más bien utilizar las siguientes diez reglas sencillas para una efectiva práctica estadística. Provienen de un grupo de veteranos estadísticos que, al igual que yo he intentado hacer en este libro, procuran enfatizar las cuestiones no técnicas que generalmente no se enseñan en los cursos de Estadística.[198] He añadido mis propios comentarios. Estas «reglas» deberían ser muy obvias, y resumir de manera eficiente las cuestiones abordadas en este libro.

1. Los métodos estadísticos deberían permitir que los datos respondan a cuestiones científicas. Pregúntese: «¿Por qué estoy haciendo esto?», en lugar de centrarse en qué técnica particular usar.

2. Las señales siempre llevan ruido. Es difícil separar las dos, y eso es lo que hace que esa materia sea interesante. La variabilidad es inevitable, y los modelos de probabilidad son una abstracción útil.

3. Planifique con antelación, con mucha antelación. Esto incluiría la idea de la preespecificación en los experimentos confirmatorios (que evita los grados de libertad del investigador).

4. Hay que preocuparse por la calidad de los datos. Todo se remite a los datos.

5. El análisis estadístico es más que un conjunto de cálculos informáticos. No se limite a dar al botón del software estadístico, tiene que saber qué es lo que está haciendo.

6. No se complique. El mensaje principal que quiera comunicar debe ser lo más básico posible; no presuma de conocimientos sobre modelos complejos a no ser que sean realmente necesarios.

7. Proporcione evaluaciones de variabilidad. Con la advertencia de que los márgenes de error son generalmente mayores de lo que se

afirma.

8. Compruebe sus supuestos. Y deje claro cuándo no ha sido posible hacerlo.

9. Siempre que pueda, ¡replique! O anime a otros a hacerlo.

10. Haga que sus análisis sean reproducibles. Otras personas deberían tener acceso a sus datos y codificaciones.

La ciencia estadística juega un papel importante en nuestras vidas, y está cambiando constantemente en respuesta a la creciente cantidad y profundidad de datos disponibles. Pero el estudio de la estadística no solo tiene un impacto en la sociedad en general, sino en los individuos en particular. Desde una perspectiva puramente personal, escribir este libro ha hecho que me dé cuenta de hasta qué punto mi vida se ha visto enriquecida por mi compromiso con la estadística. Espero que usted llegue a sentir lo mismo; si no ahora, en el futuro.

[198] R. E. Kass *et al.*, «Ten Simple Rules for Effective Statistical Practice», PLOS Computational Biology 12:6 (9 de junio de 2016), e1004961.

GLOSARIO

ajuste o estratificación: inclusión en un modelo de regresión de variables de confusión conocidas que no tienen un interés directo, sino que más bien se incluyen para permitir una comparación más equilibrada entre grupos. La intención es que los efectos estimados de las variables explicativas de interés estén, como consecuencia, más cerca de efectos causales.

alfabetización en datos: capacidad de entender los principios del aprendizaje a través de los datos, llevar a cabo análisis básicos de datos y criticar la calidad de afirmaciones basadas en datos.

algoritmo: regla o fórmula que incorpora variables y produce un resultado, como por ejemplo una predicción, una clasificación o una probabilidad.

analítica predictiva: uso de los datos para crear algoritmos que realicen predicciones.

aprendizaje automático: procedimientos para extraer algoritmos, por ejemplo para la clasificación, la predicción o el agrupamiento a partir de datos complejos.

aprendizaje no supervisado: identificación de clases basada en casos sin una pertenencia identificada, usando alguna forma de procedimiento de agrupamiento.

aprendizaje profundo: técnica de aprendizaje automático que extiende los modelos estándar de redes neuronales artificiales a planos distintos que representan diferentes niveles de abstracción. Por ejemplo, de píxeles individuales de una imagen al reconocimiento de objetos.

aprendizaje supervisado: construcción de un algoritmo de clasificación basado en casos con pertenencia confirmada a clases.

árbol de clasificación: algoritmo para clasificar en el cual las distintas características son examinadas de manera secuencial, y la respuesta indica cuál es la siguiente característica que examinar, hasta que se hace una nueva clasificación.

bayesiano: enfoque de la inferencia estadística en el cual la probabilidad se usa no solo para la incertidumbre aleatoria, sino también para la incertidumbre epistémica sobre hechos desconocidos. El teorema de Bayes se usa entonces para revisar estas creencias a la luz de la nueva evidencia.

bootstrap: forma de generar los intervalos de confianza y la distribución de pruebas estadísticas a través del re-muestreo de los datos observados en lugar de asumir un modelo probabilístico para la variable dependiente aleatoria subyacente. Una muestra básica bootstrap de un conjunto de datos x1, x2… xn es una muestra de tamaño n con reemplazo, de manera que la muestra bootstrap será extraída del conjunto original de valores, pero no generalmente en las mismas proporciones que en la base de datos original.

calibración: requisito de que las frecuencias observadas de eventos igualen a las esperadas calculadas mediante predicción probabilística. Por ejemplo, cuando a los eventos se les da una probabilidad de 0,7, deberían ocurrir aproximadamente un 70 % de las veces.

causalidad inversa: asociación entre dos variables que inicialmente parece causal, pero podría de hecho estar operando en la dirección opuesta. Por ejemplo, las personas que no beben alcohol tienden a tener peor salud que los bebedores moderados, pero esto es en parte debido a que algunos no bebedores han dejado de beber alcohol por problemas de salud.

ciego: característica de los participantes en un ensayo clínico que no saben qué tratamiento se le ha dado a un paciente, para evitar sesgo en la evaluación del resultado. Simple ciego es que los pacientes no saben qué tratamiento han recibido, doble ciego significa que los que supervisan a los pacientes no saben cuál es su tratamiento, triple ciego es cuando los tratamientos son etiquetados, por ejemplo, como A y B, y los estadísticos que analizan los datos y el comité que supervisa los resultados no saben qué corresponde al nuevo tratamiento.

ciencia de los datos: estudio y aplicación de técnicas para derivar ideas a partir de los datos, incluida la construcción de algoritmos para la predicción. La ciencia estadística tradicional forma parte de la ciencia de los datos, que también incluye un fuerte elemento de codificación y gestión de datos.

ciencia estadística: disciplina de aprender del mundo a partir de los datos, normalmente mediante un ciclo de resolución de problemas como el PPDAC.

coeficiente de correlación de Pearson: para un conjunto de n números emparejados (x1, y1), (x2, y2)… (xn, yn), donde x, sx son la media muestral y la desviación típica de las x, e y, sy son la media muestral y la desviación típica de las y, el coeficiente de correlación de Pearson viene dado por:

r = (Σni = 1 (x1 − x) (y1 − y)) / (Σni = 1 (x1 − x)2 Σni = 1 (y1 − y)2)

Supongamos que las x y las y han sido estandarizadas a puntuaciones Z dadas por u y v respectivamente, de manera que ui = (xi − x) / sx, y vi = (yi − y) / sy. Entonces, el coeficiente de correlación puede expresarse como Σni = 1 ui vi, que es el «producto cruzado» de las puntuaciones Z.

coeficiente de regresión: parámetro estimado en un modelo estadístico que expresa la fortaleza de la relación entre una variable explicativa y un resultado en el análisis de regresión múltiple. El coeficiente tendrá una interpretación distinta dependiendo de si la variable de resultado es continua (regresión lineal múltiple), un porcentaje (regresión logística), un recuento (regresión Poisson) o un tiempo de supervivencia (regresión de Cox).

comportamiento inductivo: propuesta de Jerzy Neyman y Egon Pearson en la década de 1930 para presentar la comprobación de hipótesis en el marco de la toma de decisiones. Las ideas de significatividad, potencia y errores de tipo I y tipo II son resultado de ello.

comprobación de hipótesis: procedimiento formal para evaluar el apoyo a las hipótesis que proporcionan los datos, generalmente una amalgama de pruebas clásicas fisherianas de una hipótesis nula usando un valor P y la estructura Neyman-Pearson de hipótesis nula y alternativa y errores de tipo I y de tipo II.

contracción: influencia de una distribución previa en análisis bayesiano, en la que un estimador tiende a ser arrastrado o bien hacia una media previa asumida, o bien hacia una estimada. También se conoce como «fuerza prestada», dado que, por ejemplo, las tasas estimadas de una enfermedad en un área geográfica específica se ven influidas por las tasas en otras áreas.

contrafactual: posibilidad del tipo «qué pasaría si» en la que se considera una historia de eventos alternativos.

corrección de Bonferroni: método para ajustar el tamaño (error de tipo I) de los intervalos de confianza para permitir la comprobación simultánea de múltiples hipótesis. Específicamente, cuando se comprueban n hipótesis para un tamaño global (error de tipo I) de α, cada hipótesis es comprobada con tamaño α / n. Igualmente, se asocian intervalos de confianza de 100(1 − α / n) % a cada cantidad

estimada. Por ejemplo, cuando se comprueban 10 hipótesis con un α global de 5 %, los valores P deben ser comparados para 0,05 / 10 = 0,005, y se usan intervalos de confianza de 99,5 %.

correlación de rango de Spearman: el rango de una observación es su posición en el conjunto ordenado, en el que se considera que los «empates» tienen el mismo rango. Por ejemplo, para los datos (3, 2, 1, 0, 1) los rangos son (5, 4, 2,5, 1, 2,5). La correlación de rango de Spearman es simplemente la correlación de Pearson cuando las x e y son reemplazadas por sus respectivos rangos.

crisis de reproductibilidad: afirmación de que muchos descubrimientos científicos publicados se basan en estudios de una calidad insuficiente, de manera que los resultados no son reproducibles por otros investigadores.

cuartiles (de una población): los percentiles 25, 50 y 75.

curva de característica operativa del receptor (ROC):[199] para un algoritmo que genera una puntuación, podemos escoger un umbral particular por encima del cual una unidad es clasificada como «positiva». A medida que este umbral varía, la curva ROC se forma situando la sensibilidad resultante (tasa de positivos reales) en el eje de las y frente a 1 menos la especificidad (tasa de falsos negativos) en el eje de las x.

datos binarios: variables que solo pueden adoptar dos valores, a menudo respuestas de sí o no a una pregunta. Pueden ser matemáticamente representados por una distribución de Bernoulli.

desviación típica (DT): raíz cuadrada de la varianza de una muestra o distribución. Para distribuciones de datos razonablemente simétricas sin largas colas, esperaríamos que la mayor parte de las observaciones estén dentro de dos desviaciones típicas muestrales con respecto a la media de la muestra.

distribución binomial: cuando hay n posibilidades independientes de que ocurra un evento, cada una de ellas con la misma probabilidad, el número observado de eventos tiene una distribución binomial. Técnicamente, para n ensayos independientes de Bernoulli X1, X2… Xn, cada uno de ellos con probabilidad p de éxito, su suma R = X1 + X2 + … + Xn tiene una distribución binomial con media np y varianza np (1 − p), donde P(R = r) = (nr)pr(1 − p)n−r. La proporción observada R/n tiene media p y varianza p(1 − p)/n; R/n puede, por tanto, ser considerado como un estimador de p, con error

típico $\sqrt{p(1-p)/n}$.

distribución de Bernoulli: si X es una variable aleatoria que toma el valor 1 con probabilidad p, y 0 con probabilidad $1-p$, se conoce como un ensayo de Bernoulli con una distribución de Bernoulli. X tiene media p y varianza $p(1-p)$.

distribución de probabilidad: término genérico para una expresión matemática de la probabilidad de que una variable aleatoria adopte unos valores particulares. Una variable aleatoria X tiene una función de distribución de probabilidad definida por $F(x) = P(X \leq x)$, para todo $-\infty < x < \infty$, es decir, que la probabilidad de X es como mucho x.

distribución empírica o de los datos: pauta formada por un conjunto de observaciones numéricas o categóricas.

distribución hipergeométrica: probabilidad de k éxitos en n muestreos, sin reemplazo, a partir de una población finita de tamaño N que contiene exactamente K objetos con esa característica. Formalmente estaría dada por:

$(Kk)(Nn--Kk)/(Nn)$

distribución muestral: distribución de probabilidad de un estadístico.

distribución normal: X tiene una distribución normal (gaussiana) con media μ y varianza $\sigma 2$ si tiene la siguiente función de probabilidad de densidad: $f(x) = (1/\sqrt{2\pi\sigma 2})e-(x-\mu)2 / 2\sigma 2$, para $-\infty \leq x \leq \infty$. Entonces $E(X) = \mu$, $V(X) = \sigma 2$, $DT = \sigma$. La variable estandarizada $Z = (X-\mu) / \sigma$ tiene media 0 y varianza 1, y se dice que tiene una distribución normal estándar. Empleamos Φ para referirnos a la probabilidad acumulada de una variable normal estándar Z.

Por ejemplo, $\Phi(-1) = 0{,}16$ es la probabilidad de que una variable normal estándar sea menor de -1, o, lo que es lo mismo, la probabilidad de que una variable general normal sea menor de una desviación típica por debajo de la media. El percentil mayor 100p % de la distribución normal estándar es zp, donde $P(Z \leq zp) =$

$= p$. Los valores de Φ están disponibles en el software o en tablas estándar, así como los puntos porcentuales zp: por ejemplo, el percentil 75 de la distribución normal estándar es $z0{,}75 = 0{,}67$.

distribución poblacional: cuando la población literalmente existe, pauta de observaciones potenciales para toda la población. También se

refiere a la distribución de probabilidad de una variable aleatoria genérica.

distribución Poisson: distribución para una variable aleatoria contable X para la cual $P(X = x|\mu) = e—\mu (\mu x/x!)$ para $x = 0, 1, 2\ldots$ Entonces, $E(X) = \mu$ y $V(X) = \mu$.

distribución posterior: en análisis bayesiano, distribución de probabilidad de parámetros desconocidos después de tener en cuenta los datos observados a través del teorema de Bayes.

distribución previa: en análisis bayesiano, distribución de probabilidad inicial para parámetros desconocidos. Después de observar los datos, se revisa para alcanzar una distribución posterior usando el teorema de Bayes.

distribuciones previas objetivas: intento de eliminar el elemento subjetivo del análisis bayesiano, preestableciendo distribuciones previas con la intención de representar ignorancia sobre parámetros, dejando que los datos hablen por sí mismos. No se ha establecido ningún procedimiento general para establecer esas distribuciones previas.

ensayo aleatorio controlado (EAC): diseño experimental en el que personas u otras unidades investigadas son asignadas aleatoriamente a diferentes intervenciones, asegurándose de esa manera de que los grupos están equilibrados tanto en los factores conocidos como en los desconocidos. Si los grupos muestran diferencias subsiguientes en los resultados, entonces o bien el efecto debe ser la consecuencia de la intervención, o bien ha ocurrido un evento sorprendente, cuya probabilidad puede expresarse como un valor P.

epidemiología: estudio de las tasas y las razones de ocurrencia de enfermedades.

epidemiología forense: uso del conocimiento sobre las causas de las enfermedades en las poblaciones cuando se realizan juicios sobre las causas de enfermedades en individuos.

error cuadrático medio (ECM): medida de rendimiento cuando se hacen predicciones t1…tn a partir de observaciones x1…xn, dado por $(1/n)\ \Sigma ni = 1(xi - ti)2$.

error de tipo I: hipótesis nula verdadera que es incorrectamente rechazada en favor de una alternativa, de tal manera que se hace una afirmación de falso positivo.

error de tipo II: hipótesis alternativa que es cierta, pero una prueba de hipótesis no rechaza la hipótesis nula, de manera que la conclusión es un falso negativo.

error residual: término genérico para el componente de los datos que no puede ser explicado por un modelo estadístico, por lo que se dice que se debe a variación aleatoria.

error típico: desviación típica de una media muestral, cuando se considera como una variable aleatoria. Supongamos que X1, X2… Xn son variables independientes idénticamente distribuidas extraídas de una distribución poblacional de media μ y desviación típica σ. Entonces su promedio Y = (X1 + X2 + … +Xn)/n tiene media μ y varianza σ2/n. La desviación típica de Y es σ/√n, también conocida como error típico, y estimado por s/√n, donde s es la desviación típica de la muestra de las X observadas.

escala logarítmica: el logaritmo en base 10 de un número positivo x es y = log10 x, o bien x = 10y. En análisis estadístico, log x normalmente se refiere al logaritmo natural y = loge x, o bien x = ey, siendo e la constante exponencial 2,718.

especificidad: proporción de casos «negativos» que son correctamente identificados por medio de un clasificador o una prueba. Uno menos la especificidad es también conocido como el error observado de tipo I, o tasa de falso positivo.

esperanza (media): promedio-media de una variable aleatoria. Se define como Σxp(x) para una variable aleatoria discreta X y ∫xp(x)dx para una variable aleatoria continua. Por ejemplo, si X es el resultado de lanzar un dado justo, entonces P(X = x) = 1/6 para x = 1, 2, 3, 4, 5, 6, de manera que E(X) = 1/6 (1 + 2 + 3 + 4 + 5 + 6) = 3,5.

estadístico: número coherente que se deriva de un conjunto de datos.

estadístico-t: prueba estadística usada para comprobar una hipótesis nula de que un parámetro es 0, formada por la razón de un estimador con respecto a su error típico. Para grandes muestras, valores por encima de 2 o por debajo de −2 corresponden a un valor P de dos colas de 0,05; pueden obtenerse valores P exactos mediante programas estadísticos.

estudio de caso y control: diseño de estudio retrospectivo en el que individuos con una enfermedad o un resultado de interés (los casos) son emparejados con una o más personas que no tienen la enfermedad (los controles), y las historias de los dos grupos son comparadas para

ver si hay situaciones que difieren sistemáticamente entre los dos grupos. Este diseño solo puede estimar riesgos relativos asociados a situaciones.

estudio prospectivo de cohorte: estudio en el que un conjunto de individuos son identificados, se miden sus factores contextuales y después se los sigue en el tiempo y se observan resultados relevantes. Esos estudios son largos y caros, y podrían no identificar muchos eventos poco frecuentes.

estudio retrospectivo de cohorte: estudio en el que un conjunto de individuos son identificados en un punto del pasado y se siguen en el tiempo sus resultados subsiguientes hasta la actualidad. Un estudio de ese tipo no requiere un período amplio de seguimiento, pero depende de que las variables explicativas apropiadas hayan sido medidas en el pasado.

estudio transversal: análisis que se basa únicamente en el estado actual de los individuos, sin seguimiento en el tiempo.

estudios confirmatorios: estudios rigurosos que idealmente se realizan siguiendo un protocolo preespecificado para confirmar o negar hipótesis sugeridas por estudios exploratorios.

estudios exploratorios: estudios iniciales flexibles que permiten cambios adaptativos en el diseño y los análisis para seguir indicios prometedores, y tienen como objetivo generar hipótesis que sean comprobadas en estudios confirmatorios.

eventos dependientes: situación en la que la probabilidad de un evento depende del resultado de otro evento.

eventos independientes: A y B son independientes si la ocurrencia de A no influye en la probabilidad de B, tal que $p(B/A) = p(B)$, o, lo que es igual, $p(B, A) = p(B)p(A)$.

exposición: factor del que interesa evaluar su impacto en una enfermedad, muerte y otro resultado médico, como por ejemplo un aspecto del entorno o un comportamiento.

factor de Bayes: apoyo relativo que un conjunto de datos proporciona a dos hipótesis alternativas. Para las hipótesis H0 y H1, y los datos x, la razón es $p(x|H0)/p(x|H1)$.

factor oculto: en epidemiología, exposición que no ha sido medida pero puede ser una variable de confusión responsable de algunas de

las asociaciones observadas: por ejemplo, cuando el estatus socioeconómico no ha sido medido en un estudio sobre la relación entre la dieta y la enfermedad.

falacia de la evidencia incompleta: cuando la probabilidad de que se incluya a una persona en la muestra, o se observe una característica, depende de algún factor contextual, por ejemplo cuando se supervisa más estrechamente a los asignados al grupo de tratamiento en un ensayo aleatorizado que al grupo de control.

falacia del fiscal: pequeña probabilidad sobre una evidencia, dado que el acusado es inocente, que es interpretada erróneamente como la probabilidad de la inocencia, dada la evidencia.

falso positivo: clasificación incorrecta de un caso «negativo» como «positivo».

frecuencias esperadas: número de eventos que se espera que ocurran en el futuro, de acuerdo con un modelo de probabilidad asumido.

gráfico de embudo: gráfico de un conjunto de observaciones de diferentes medidas cruzado con una medida de su precisión, y en el que las unidades pueden ser instituciones, áreas o estudios. A menudo, dos «embudos» indican dónde esperaríamos que se encontrasen el 95 % y el 99,8 % de las observaciones, en el caso de que no hubiese realmente diferencias subyacentes entre las unidades. Cuando la distribución de las observaciones es aproximadamente normal, los límites de control de 95 % y 99,8 % son esencialmente la media ± dos y tres errores típicos.

grupo de control: grupo de individuos que no han sido sometidos a la exposición de interés, mediante aleatorización.

hipótesis nula: teoría científica por defecto, que representa generalmente la ausencia de un efecto o un descubrimiento de interés, que es comprobada usando un valor P. Generalmente se escribe H0.

incertidumbre aleatoria: impredecibilidad inevitable acerca del futuro, también llamada casualidad, azar y suerte, entre otros.

incertidumbre epistémica: desconocimiento sobre hechos, números o hipótesis científicas.

inferencia estadística: proceso de usar datos muestrales para aprender sobre parámetros desconocidos subyacentes a un modelo estadístico.

inferencia inductiva (inducción): proceso de aprender sobre principios generales a partir de ejemplos específicos.

ingeniería de características: en el campo del aprendizaje automático, proceso de reducir la dimensionalidad de las variables de entrada, creando indicadores resumidos con la intención de abarcar la información contenida en la totalidad de los datos.

inteligencia artificial (IA): conjunto de programas informáticos creados con la intención de desarrollar tareas normalmente asociadas con capacidades humanas.

intención de tratar: principio por el cual los participantes en ensayos aleatorizados son analizados de acuerdo con la intervención que se supone que van a recibir, la hayan recibido realmente o no.

interacciones: múltiples variables explicativas que se combinan para producir un efecto diferente al esperado de sus contribuciones individuales.

intervalo de confianza: intervalo estimado dentro del cual puede plausiblemente encontrarse un parámetro desconocido. Para un conjunto observado de datos x, un intervalo de confianza del 95 % para μ es un intervalo cuyo límite inferior L(x) y límite superior U(x) tiene la propiedad de que, antes de observar los datos, hay una probabilidad del 95 % de que el intervalo aleatorio (L(X), U(X)) contenga μ. El teorema central del límite, combinado con el conocimiento de que cerca del 95 % de una distribución normal se encuentra entre la media ± 2 desviaciones típicas, significa que una aproximación común a un intervalo de confianza del 95 % es ± 2 errores típicos esperados. Supongamos que queremos encontrar un intervalo de confianza para la diferencia $\mu 2 - \mu 1$ entre los dos parámetros $\mu 2$ y $\mu 1$. Si T1 es un estimador de $\mu 1$ con error típico ET2, entonces $T2 - T1$ es un estimador de $\mu 2 - \mu 1$. La varianza de la diferencia entre los dos estimadores es la suma de sus varianzas, de manera que el error típico de $T2 - T1$ está dado por $\sqrt{SE12 + SE22}$. A partir de ahí, se puede construir un intervalo de confianza $\mu 2 - \mu 1$.

ley de los números grandes: proceso por el cual la media muestral de un conjunto de variables aleatorias tiende hacia la media de la población.

límites de control: límites preespecificados para una variable aleatoria que se usan en control de calidad para supervisar la desviación con respecto a un determinado estándar, y que, por ejemplo, se pueden

presentar en un gráfico de embudo.

macrodatos: término crecientemente anacrónico caracterizado a veces por cuatro uves: un enorme Volumen de datos, una Variedad de fuentes tales como imágenes, cuentas en redes sociales o transacciones, una alta Velocidad en su adquisición y una posible falta de Veracidad debido a su rutinaria recolección.

marco: forma de presentar las cifras que, a su vez, puede influir en la impresión causada en el público.

margen de error: en una encuesta, rango posible en el cual puede encontrarse una característica real de una población. Son normalmente intervalos de confianza del 95 %, que son aproximadamente ± 2 errores típicos, pero a menudo se emplean barras de error para representar ± 1 error típico.

matriz de errores: tabla de contingencia de la clasificación correcta e incorrecta realizada por medio de un algoritmo.

matriz de iconos: presentación gráfica de frecuencias a través de un conjunto de pequeñas imágenes, por ejemplo de personas.

media (de una muestra): supongamos que tenemos un conjunto de n puntos de datos, que denominamos como x1, x2, …, xn. Entonces, su media muestral viene dada por m = (x1 + x2 + …+ xn)/n, que puede ser escrito como m = ½ Σni=1 xi = x. Por ejemplo, si 3, 2, 1, 0, 1 son el número de niños que tienen 5 personas en una muestra, entonces la media muestral es (3 + 2 + 1 + 0 + 1)/5 = 7/5 = 1,4.

media (de una población): véase esperanza.

media muestral: véase media (de una muestra).

mediana (de una muestra): valor que está en el medio de un conjunto ordenado de puntos de datos. Si los puntos de datos se ponen en orden, denominamos al más bajo x(1), al segundo más bajo x(2), y así sucesivamente hasta el valor máximo x(n). Si n es impar, entonces la mediana de la muestra es el valor que está en el medio x(n + 1)/2; si n es par, entonces se toma como mediana la media de los dos puntos «medios».

metaanálisis: método estadístico formal para combinar los resultados de múltiples estudios.

mínimos cuadrados: supongamos que tenemos un conjunto de n números emparejados (x1, y1), (x2, y2)…, (xn, yn), y x, sx son la media muestral y la desviación típica de las y. Entonces, la regresión de mínimos cuadrados viene dada por:

y = b0 + b1 (x − x), donde:

→ y es el valor predicho de la variable dependiente y para un valor específico de la variable independiente x.

→ La pendiente es b1 = (Σi (yi − y)(xi − x))/(Σi (xi − x)2).

→ La constante es b0 = y. La línea de mínimos cuadrados pasa por el centro de gravedad x, y.

→ El residuo i-ésimo es la diferencia entre la observación i-ésima y su valor predicho, yi − yi.

→ El valor ajustado de la observación i-ésima es el residuo sumado a la constante, es decir yi − yi + y. Se supone que debe ser el valor que habríamos observado si se tratase de un caso «promedio», es decir, con x = x, en lugar de x = xi.

→ La suma del cuadrado de los residuos (SCR) es la suma de los cuadrados de los residuos tal que SCR = Σni=1 (yi − y)2.

→ La pendiente b1 y el coeficiente de correlación de Pearson r se relacionan a través de la fórmula b1 = rsy /sx. De manera que si las desviaciones típicas de las x y las y son las mismas, entonces la pendiente es exactamente igual al coeficiente de correlación.

moda (de una distribución poblacional): respuesta que tiene la mayor probabilidad de ocurrencia.

moda (de una muestra): valor más común de un conjunto de datos.

modelo estadístico: representación matemática, que contiene parámetros desconocidos, de la distribución de probabilidad de un conjunto de variables aleatorias.

modelos jerárquicos: en análisis bayesiano, cuando se asume que los parámetros subyacentes a un número de unidades, por ejemplo áreas o colegios, han sido extraídos de una distribución previa común. Esto tiene como resultado que los parámetros estimados de unidades individuales converjan hacia una media global.

nivel de significatividad: tasa de error de tipo I de un test estadístico, generalmente denominada α.

paradoja de Simpson: relación aparente que cambia su signo cuando se tiene en cuenta una variable de confusión.

parámetros: aquellas cantidades desconocidas en un modelo estadístico, generalmente indicadas mediante letras griegas.

percentil (de una muestra): el percentil 70 de una muestra, por ejemplo, es el valor que está en el 70 % del conjunto de datos ordenado; la mediana es, por tanto, el percentil 50. Podría ser

necesario interpolar entre puntos.

percentil (de una población): hay, por ejemplo, una probabilidad del 70 % de extraer una observación aleatoria por debajo del percentil 70. Para una población literal, es el valor por debajo del cual se encuentra el 70 % de la población.

placebo: tratamiento falso aplicado al grupo de control de un ensayo clínico aleatorizado, como por ejemplo una píldora de azúcar parecida al tratamiento que se está probando.

población: grupo del cual se asume que se extraen los datos muestrales, y que proporciona la distribución de probabilidad de una única observación. En una encuesta esta podría ser una población literal, pero cuando se realizan mediciones, o cuando se tienen todos los datos posibles, la población se convierte en una idealización matemática.

potencia: probabilidad de rechazar correctamente la hipótesis nula, dado que la hipótesis alternativa es cierta. Es 1 menos la tasa de error de tipo II de una prueba estadística, y generalmente se escribe $1 - \beta$.

PPDAC: propuesta de «ciclo de datos» consistente en Problema, Plan, reunión de Datos, Análisis (exploratorio o confirmatorio) y Conclusiones y comunicación.

predicción probabilística: predicción en forma de distribución de probabilidad de un evento futuro, más que un juicio categórico de lo que ocurrirá.

probabilidad: expresión matemática formal de la incertidumbre. Sea P(A) la probabilidad de un evento A. Entonces, las reglas de la probabilidad son:

1. Límites: $0 \leq P(A) \leq 1$, con $P(A) = 0$ si A es imposible y $P(A) = 1$ si A es segura.

2. Complemento: $P(A) = 1 - P(\text{NO A})$.

3. Regla de la suma: Si A y B son mutuamente exclusivas (es decir, que como mucho puede ocurrir una de ellas), $P(\text{A O B}) = P(A) + P(B)$.

4. Regla de la multiplicación: Para todo evento A y B, $P(\text{A Y B}) = P(A|B)P(B)$, donde $P(A|B)$ representa la probabilidad de A dado que ha ocurrido B. A y B son independientes si y solo si $P(A/B) = P(A)$, es decir, que la ocurrencia de B no afecta a la probabilidad de A. En este

caso tenemos P(A Y B) = P(A)P(B), la regla de multiplicación para eventos independientes.

probabilidad de coincidencia aleatoria: en pruebas forenses de ADN, probabilidad de que una persona seleccionada aleatoriamente de una población relevante coincida con el perfil de ADN observado que conecta a un sospechoso con un crimen.

promedio: término genérico referido a un único valor representativo para un conjunto de números, por ejemplo la media, la mediana o la moda.

prueba secuencial: prueba estadística que es llevada a cabo repetidamente para datos acumulados, inflando así la probabilidad de que se produzca en algún momento un error de tipo I. Se garantiza un «resultado significativo» si se sigue con este proceso el tiempo suficiente.

pruebas múltiples: implementación de una serie de pruebas de hipótesis, incrementando la probabilidad de un falso positivo (error de tipo I).

pruebas unilateral y bilateral: la prueba de hipótesis unilateral se usa cuando una hipótesis nula especifica que, por ejemplo, el efecto de un tratamiento médico es negativo. Esto solo puede ser rechazado por grandes valores positivos de una prueba estadística que represente un efecto de tratamiento estimado. Una prueba bilateral sería apropiada para una hipótesis nula de que un efecto de tratamiento fuese, digamos, exactamente cero, de manera que tanto unas estimaciones positivas como unas negativas llevarían a rechazar la hipótesis nula.

puntuación de Brier: medida de la precisión de las predicciones probabilísticas, basada en la media del error predicho al cuadrado. Si p1…pn son las probabilidades dadas a que un conjunto de n observaciones binarias x1…xn tomen los valores 0 y 1, entonces la puntuación de Brier es (1/n) (Σni (xi − pi)2). Se trata básicamente de un criterio de error medio al cuadrado aplicado a datos binarios.

puntuación Z: forma de estandarizar una observación xi en términos de su distancia con respecto a la media muestral m expresada en términos de las desviaciones típicas de la muestra s, de manera que zi = (xi − m)/s. Una observación con una puntuación Z de 3 corresponde a estar 3 desviaciones típicas por encima de la media, que es un valor típico considerablemente extremo. Una puntuación Z también puede ser definida en términos de una media poblacional μ y

una desviación típica σ, en cuyo caso zi = (xi − μ)/σ.

rango (de una muestra): el máximo menos el mínimo, escrito como x(n) − x(1).

rango intercuartílico: medida de la dispersión de una muestra o de una distribución poblacional, específicamente la distancia entre los percentiles 25 y 75. Equivalente a la diferencia entre el primer y el tercer cuartil.

razón de probabilidad: si la probabilidad de un evento es p, la razón de probabilidad de un evento se define como p/(1 − p). Si la razón de probabilidad de un evento en el grupo expuesto es p/(1 − p), y la razón de probabilidad del grupo no expuesto es q/(1 − q), la razón de probabilidad viene dada entonces por (p/(1 − p))/(q/(1 − q)). Si p y q son pequeñas, entonces la razón de probabilidad estará cerca del riesgo relativo p/q, pero las razones de probabilidad y los riesgos relativos comienzan a diferir cuando los riesgos absolutos son mucho mayores del 20 %.

razón de riesgo: al analizar tiempos de supervivencia, riesgo relativo (asociado a una exposición) de sufrir un evento en un período fijo de tiempo. Una regresión de Cox es una forma de regresión múltiple en la que la variable de respuesta es el tiempo de supervivencia, y los coeficientes corresponden a log(razones de riesgo).

razón de tasas de incidencia: incremento relativo en el número esperado de eventos en un período fijo de tiempo asociado con una exposición. Una regresión Poisson es una forma de regresión múltiple en la que la variable dependiente es la tasa observada, y los coeficientes corresponden a log(razones de tasas).

razón de verosimilitud: medida del apoyo relativo que unos datos proporcionan a dos hipótesis rivales. Para las hipótesis H0 y H1, la razón de verosimilitud proporcionada por los datos x viene dada por p(x|H0)/p(x|H1).

regresión a la media: observación alta o baja que es seguida por otra que es menos extrema por un proceso de variación natural. Ocurre porque parte de la explicación del caso extremo inicial es el azar, y es poco probable que se repita en la misma medida.

regresión de Cox: véase razón de riesgo.

regresión lineal múltiple: supongamos que para cada respuesta yi hay un conjunto p de variables independientes (xi1, xi2… xip). Entonces,

una regresión lineal múltiple de mínimos cuadrados viene dada por:

$$yi = b0 + b1\,(xi1 - x1) + b2\,(xi2 - x2) + \ldots + bp\,(xip - xp),$$

donde los coeficientes b0, b1… bp son escogidos para minimizar la suma cuadrática residual $SCR = \Sigma ni=1\,(yi - yi)2$. La constante b0 es simplemente la media y, mientras que la fórmula para los demás coeficientes es compleja pero fácil de calcular. Nótese que $b0 = y$ es el valor predicho de una observación y cuyas variables independientes son las medias (x1, x2… xp), y al igual que en el caso de la regresión lineal, se obtiene una yi ajustada a partir de los residuos más la constante, o $yi - yi + y$.

regresión logística: forma de regresión múltiple en la que la variable de respuesta es una proporción, y los coeficientes corresponden a log(razones de probabilidad). Supongamos que observamos una serie de proporciones $yi = ri/ni$, que se asume que se derivan de una variable binomial con una probabilidad subyacente pi con un correspondiente conjunto de variables independientes (xi1, xi2… xip). Se asume que el logaritmo de la razón de probabilidad pi es una regresión lineal:

$$\log (pi\,/(1 - pi)) = b0 + b1xi1 + b2xi2 + \ldots + bpxip$$

Supongamos que una de las variables independientes, por ejemplo x1, es binaria con $x1 = 0$, lo que se corresponde a no estar expuesta a un riesgo potencial, y $x1 = 1$, que corresponde a estar expuesta. Entonces el coeficiente b1 es un log(razón de probabilidad).

regresión multinivel y post-estratificación (RMP): moderno desarrollo en el campo de las muestras basadas en encuestas en el cual se obtiene un número singularmente pequeño de entrevistados de muchas áreas. Después se construye un modelo de regresión que relaciona las respuestas con factores demográficos, controlando por variabilidad adicional entre áreas mediante modelos jerárquicos. Si se conocen los datos demográficos de todas las áreas, se pueden hacer predicciones locales y nacionales, con la incertidumbre apropiada.

residuo: diferencia entre un valor observado y el predicho por un modelo estadístico.

riesgo absoluto: porcentaje de personas en un grupo determinado que experimenta un evento de interés en un período de tiempo específico.

riesgo relativo: si el riesgo absoluto entre personas que están expuestas a algo que nos interesa es p, y el riesgo absoluto entre

personas que no están expuestas es q, entonces el riesgo relativo es p/q.

sabiduría de las multitudes: idea de que un resumen que se deriva de la opinión de un grupo está más cerca de la verdad que la opinión de la mayoría de los individuos.

sensibilidad: proporción de casos «positivos» que son correctamente identificados por un clasificador o prueba, a menudo denominada como tasa de verdaderos positivos. Uno menos la sensibilidad también se conoce como error observado de tipo II o tasa de falsos negativos.

señal y ruido: idea de que los datos observados surgen de dos componentes: una señal determinista que es en lo que estamos realmente interesados, y un ruido aleatorio que comprende el error residual. El desafío de la inferencia estadística es identificar apropiadamente ambos, y no pensar que el ruido es en realidad una señal.

significatividad estadística: se considera que un efecto observado es estadísticamente significativo cuando su valor P correspondiente a una hipótesis nula es menor de algún nivel preespecificado, digamos 0,05 o 0,001, lo que significa que un resultado tan extremo es poco probable que ocurra si la hipótesis nula, y todas las otras asunciones de los modelos, se cumplen.

significatividad práctica: propiedad del descubrimiento que tiene una importancia real. Hay estudios que pueden dar lugar a resultados estadísticamente significativos, pero sin significatividad práctica.

sobreajuste: elaboración de un modelo estadístico que está sobreadaptado al entrenamiento de los datos, de manera que su capacidad predictiva comienza a disminuir.

solución de compromiso sesgo-varianza: cuando se ajusta un modelo para predecir, una creciente complejidad llevará finalmente a un modelo que tiene menos sesgo, en el sentido de más potencial de adaptarse a los detalles del proceso subyacente, pero más varianza, dado que no hay suficientes datos para confiar en los parámetros del modelo. Tiene que llegarse a un compromiso entre estos elementos para evitar el sobreajuste.

tasa de descubrimientos falsos: al comprobar múltiples hipótesis, porcentaje de afirmaciones positivas que resultan ser falsos positivos.

teorema central del límite: tendencia de que la media de la muestra de

un conjunto de variables aleatorias tenga una distribución muestral normal, con independencia (con ciertas excepciones) de la forma de la distribución muestral subyacente de la variable aleatoria. Si cada n observaciones independientes tienen media μ y varianza σ2, entonces bajo unas asunciones amplias, su media muestral es un estimador de μ, y tiene una distribución aproximadamente normal con media μ, varianza σ2/n, y desviación típica σ/√n (también conocida como error típico del estimador).

teorema de Bayes: regla de probabilidad que muestra cómo la evidencia A actualiza las creencias previas de una proposición B para producir creencias posteriores p(B|A), a través de la fórmula p(B|A) = (p(A|B)p(B))/p(A). Esto es fácil de demostrar: dado que p(B Y A) = p(A Y B), la regla de probabilidad de la multiplicación significa que p(B|A)p(A) = P(A|B)p(B), y, dividiendo cada lado por p(A), nos da el teorema.

test de asociación o bondad de ajuste chi-cuadrado: prueba estadística que indica el grado de incompatibilidad de los datos con un modelo estadístico supuesto que comprende la hipótesis nula, que puede indicar ausencia de asociación, u otra forma matemática. Concretamente, la prueba compara un conjunto m de recuentos esperados o1, o2… om con un conjunto de valores esperados e1, e2… em que han sido calculados bajo la hipótesis nula. La versión más sencilla de la prueba estadística es la siguiente:

$X2 = \Sigma mj=1 ((oj - ej)2/ej)$

Bajo la hipótesis nula X2 tendrá aproximadamente una distribución muestral chi-cuadrada, lo que permite calcular un valor P asociado.

test de permutación-aleatorización: forma de prueba de hipótesis en la cual la distribución del estadístico de prueba bajo la hipótesis nula se obtiene permutando las etiquetas de los datos, en lugar de por medio de un detallado modelo estadístico para las variables aleatorias. Supongamos que la hipótesis nula es que una «etiqueta», como por ejemplo, ser hombre o mujer, no está asociada con un resultado. Las pruebas de aleatorización examinan todas las formas posibles por las cuales pueden ser reasignadas las etiquetas para puntos de datos individuales, cada uno de los cuales es igualmente probable bajo la hipótesis nula. Se calcula la prueba estadística para cada una de estas permutaciones, y el valor P viene dado por la proporción que lleva a pruebas estadísticas más extremas que las realmente observadas.

validación cruzada: forma de evaluar la calidad de un algoritmo usado

para la predicción o la clasificación eliminando sistemáticamente algunos casos para que sirvan de conjunto de entrenamiento.

validez externa: cuando las conclusiones de un estudio son generalizables a un grupo objetivo, más amplio que la población inmediata que ha sido estudiada. Se refiere a la relevancia de un estudio.

validez interna: cuando las conclusiones de un estudio se aplican realmente a la población de un estudio. Esto se refiere al rigor con el cual se ha realizado un estudio.

valor P: medida de la discrepancia entre los datos y la hipótesis nula. Para una hipótesis nula H0, sea T el estadístico para el cual unos valores grandes indican inconsistencia con H0. Supongamos que observamos un valor-t. Entonces, un valor P (de una sola cola) es la probabilidad de observar un valor tan extremo, si H0 fuese cierta, es decir $P(T \geq t|H0)$. Si tanto valores pequeños como grandes de T indican inconsistencia con H0, entonces el valor P de dos colas es la probabilidad de observar un valor tan grande en ambas direcciones. A menudo se considera que el valor P de dos colas es simplemente el doble del valor P de una sola cola, mientras que el programa R usa la probabilidad total de eventos que tienen una probabilidad menor de ocurrencia que el realmente observado.

valores P de una y de dos colas: los correspondientes a pruebas unilaterales y bilaterales.

variabilidad: diferencias inevitables que se producen entre las mediciones u observaciones, algunas de las cuales pueden ser explicadas por factores conocidos, mientras que las restantes son atribuidas a ruido aleatorio.

variable aleatoria: cantidad que se asume que tiene una distribución de probabilidad. Antes de ser observadas, a las variables aleatorias se las nombra mediante una letra mayúscula X, mientras que los valores observados se denominan x.

variable categórica: variable que puede tomar dos o más valores discretos, que pueden o no estar ordenados.

variable contable: variable que puede tener valores enteros 0, 1, 2, y así sucesivamente.

variable continua: variable aleatoria X que puede, al menos en principio, adoptar cualquier valor dentro de un rango específico. Tiene

una función de probabilidad de densidad f tal que $P(X \leq x) = \int x - \infty$ f(t)dt, y una esperanza dada por $E(X) = \int \infty - \infty$ x f(x)dx. La probabilidad de que X esté en el intervalo (A, B) puede calcularse usando ∫AB f(X)dx.

variable de confusión: variable que está asociada tanto con la dependiente como con el predictor, y que puede explicar parte de su aparente relación. Por ejemplo, el peso y la estatura de los niños están marcadamente correlacionados, pero gran parte de esa asociación se explica por la edad del niño.

variable dependiente, de respuesta o de resultado: variable de interés principal que deseamos predecir o explicar.

variable independiente (predictor): variable que está fija por diseño u observación, y cuya asociación con una variable de resultado podría ser de interés.

varianza: para una muestra x1… xn con media x, esto se define generalmente como $s2 = (1/(n - 1)) \Sigma ni = 1 (xi - x)2$ (aunque el denominador puede también ser n en lugar de n − 1). Para una variable aleatoria X con media μ, la varianza es $V(X) = E(X - \mu)2$. La desviación típica es la raíz cuadrada de la varianza, $DT(X) = \sqrt{V(X)}$.

verosimilitud: medida del apoyo que los datos otorgan a valores particulares de los parámetros. Cuando una distribución de probabilidad para una variable aleatoria depende de un parámetro, por ejemplo θ, entonces después de observar el dato x la verosimilitud de θ es proporcional a p(x|θ).

[199] Por sus siglas en inglés, Receiver Operating Characteristic. (N. del T.).

Índice

Portada

El arte de la estadística

Lista de figuras

Lista de tablas

Agradecimientos

El arte de la estadística

Introducción

01. Cómo entender las cosas como proporciones: datos categóricos y porcentajes

02. Cómo resumir y comunicar los números. Muchísimos números

03. Pero, en realidad, ¿por qué nos interesan los datos? Poblaciones y mediciones

04. ¿Qué causa qué?

05. Modelizar relaciones mediante la regresión

06. Algoritmos, analítica y predicción

07. ¿Qué seguridad nos merecen los resultados? Estimadores e intervalos

08. Probabilidad: el lenguaje de la incertidumbre y la variabilidad

09. Juntar la probabilidad y la estadística

10. Responder preguntas y reivindicar descubrimientos

11. Aprender de la experiencia a la manera bayesiana

12. Cuando las cosas se hacen mal

13. Cómo podemos hacer mejor los análisis estadísticos

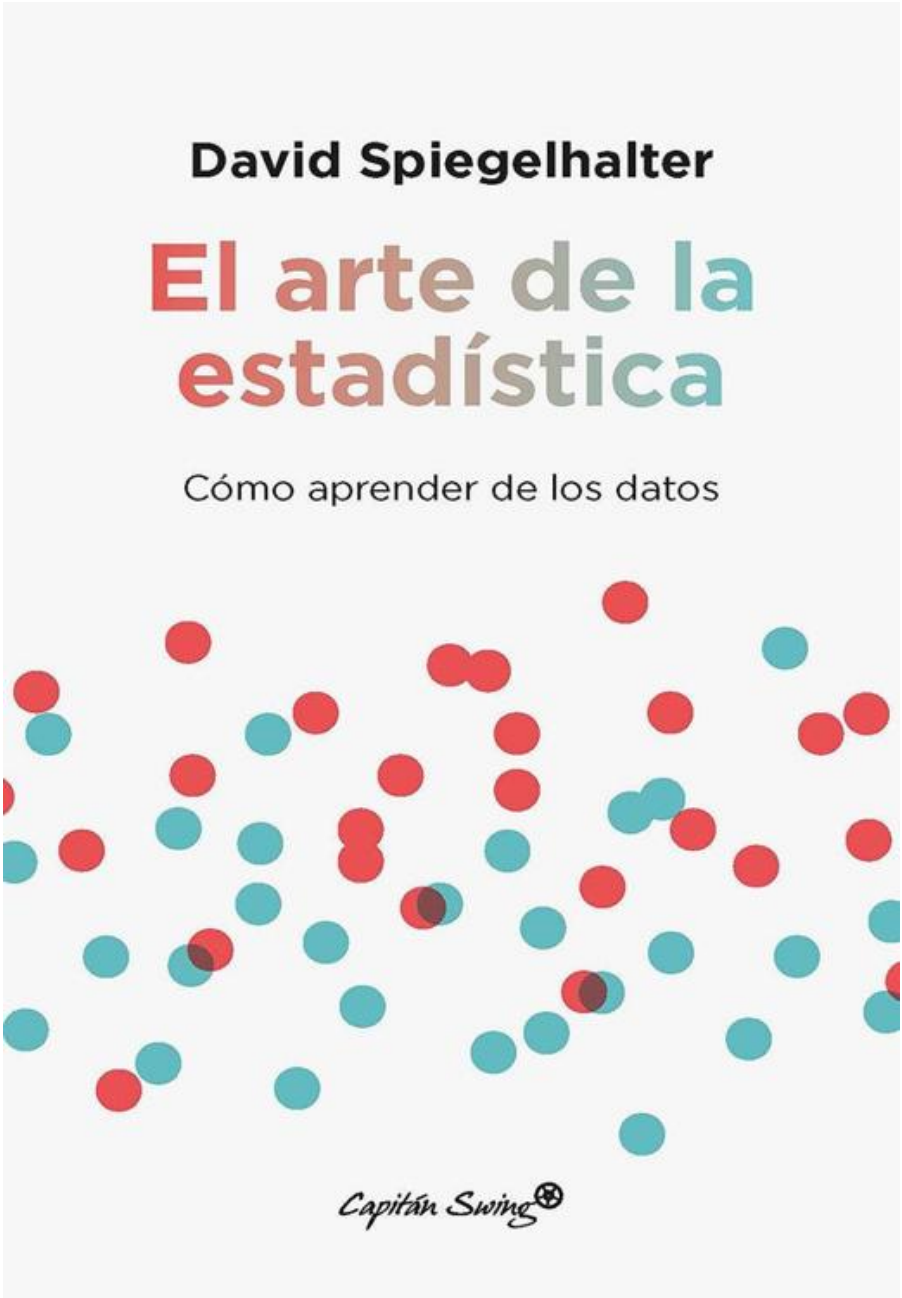
14. En conclusión

Glosario

Sobre este libro

Sobre David Spiegelhalter

Créditos



La guía definitiva del pensamiento estadístico.
Las estadísticas están en todas partes, tanto en la ciencia como en los negocios, y en los medios de comunicación populares cientos de veces

al día. En esta era de los grandes datos, un conocimiento básico de la estadística es más importante que nunca si queremos separar los hechos de la ficción, los adornos ostentosos de las pruebas en bruto, y más aún si esperamos participar en el futuro, en lugar de ser simples espectadores.

En 'El arte de la estadística', el reconocido estadístico David Spiegelhalter muestra a los oyentes cómo extraer conocimientos de los datos en bruto centrándose en los conceptos y las conexiones que hay detrás de las matemáticas. Basándose en ejemplos del mundo real para introducir cuestiones complejas, nos muestra cómo la estadística puede ayudarnos a determinar cuál fue el pasajero más afortunado del Titanic, si un famoso asesino en serie podría haber sido capturado antes y si el cribado del cáncer de ovarios es beneficioso.

'El arte de la estadística' no sólo nos muestra cómo los matemáticos han utilizado la ciencia estadística para resolver estos problemas, sino que también nos enseña cómo podemos pensar como estadísticos. Aprendemos a aclarar nuestras preguntas, suposiciones y expectativas al abordar un problema, y -quizás aún más importante-aprendemos a interpretar responsablemente las respuestas que recibimos.

David Spiegelhalter. Estadístico británico y miembro del Churchill College, Cambridge. De 2007 a 2018 fue Profesor Winton de Comprensión Pública del Riesgo en el Laboratorio de Estadística de la Universidad de Cambridge. Spiegelhalter es un investigador altamente citado por el ISI. Es presidente del Centro Winton de Comunicación de Riesgos y Pruebas en el Centro de Ciencias Matemáticas de Cambridge. El 27 de mayo de 2020 se incorporó al consejo de la Autoridad Estadística del Reino Unido como director no ejecutivo por un periodo de tres años. En 2012 Spiegelhalter presentó el documental de la BBC *Four Tails You Win: The Science of Chance*, que describía la aplicación de la probabilidad en la vida cotidiana. También presentó una charla en el Festival de Ciencia de Cambridge de 2013, *How to Spot a Shabby Statistic*, en el Babbage Lecture Theatre de Cambridge. Fue elegido presidente de la Royal Statistical Society, y asumió el cargo el 1 de enero de 2017.En marzo de 2020 Spiegelhalter lanzó un podcast llamado *Risky Talk* en el que entrevista a expertos en comunicación de riesgos y pruebas sobre temas como la genética, la nutrición, el cambio climático y la inmigración.

Título original: *The Art of Statistics: Learning from Data* (2020)

© Del libro: David Spiegelhalter

© De la traducción: Francisco Herreros

Edición en ebook: mayo de 2023

© Capitán Swing Libros, S. L.

c/ Rafael Finat 58, 2º 4 - 28044 Madrid

Tlf: (+34) 630 022 531

28044 Madrid (España)

contacto@capitanswing.com

www.capitanswing.com

ISBN: 978-84-12687-81-1

Diseño de colección: Filo Estudio - www.filoestudio.com

Corrección ortotipográfica: Victoria Parra Ortiz

Composición digital: leerendigital.com

Cualquier forma de reproducción, distribución, comunicación pública o transformación de esta obra solo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Diríjase a CEDRO (Centro Español de Derechos Reprográficos, www.cedro.org) si necesita fotocopiar o escanear algún fragmento de

esta obra.